What is 'Firm Heterogeneity' in Trade Models? The Role of Quality, Scope, Markups, and Cost*

Colin Hottman
Columbia University†

Stephen J. Redding
Princeton University and NBER‡

David E. Weinstein
Columbia University and NBER§

September 23, 2014

Abstract

We develop and structurally estimate a model of heterogeneous multiproduct firms that can be used to decompose the firm-size distribution into the contributions of costs, quality, markups, and product scope. Using Nielsen barcode data on prices and sales, we find that variation in firm quality and product scope explains at least four fifths of the variation in firm sales. We show that the imperfect substitutability of products within firms, and the fact that larger firms supply more products than smaller firms, implies that standard productivity measures are not independent of demand system assumptions and probably dramatically understate the relative productivity of the largest firms. Although most firms are well approximated by the monopolistic competition benchmark of constant markups, we find that the largest firms that account for most of aggregate sales depart substantially from this benchmark, and exhibit both variable markups and substantial cannibalization effects.

JEL CLASSIFICATION: L11, L21, L25, L60

KEYWORDS: firm heterogeneity, multiproduct firms, cannibalization effects

*We are grateful to Andrew Bernard, Swati Dhingra, Cecilia Fieler, Oleg Itskhoki, Bernard Salanié and conference and seminar participants at ERWIT, Dartmouth and Princeton for helpful comments. Thanks to Ildiko Magyari for providing excellent research assistance. David Weinstein would like to thank the NSF (Award 1127493) for generous financial support. We also would like to thank GS1 for providing us with a concordance between the barcode data and firm identifiers. The usual disclaimer applies. All results are calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

†420 W. 118th Street, MC 3308, New York, NY 10027. Email: cjh2164@columbia.edu.

‡Fisher Hall, Princeton, NJ 08544. Email: reddings@princeton.edu.

§420 W. 118th Street, MC 3308, New York, NY 10027. Email: dew35@columbia.edu.

1 Introduction

Why are some firms larger than others? Some companies, such as the Coca Cola Corporation, generate billions of dollars of sales and dominate the markets in which they operate. Other companies account for only a small fraction of the sales of their larger competitors. What explains these vast differences in firm performance? Answering this question is important for quantifying equilibrium models of firm heterogeneity developed in the recent trade and macro literatures and for understanding the relationship between microeconomic firm performance and macroeconomic outcomes.

Recent research on firm heterogeneity in trade and macroeconomics (e.g. Melitz 2003, Feenstra 2014, Manova and Zhang 2012) points to four components of firm heterogeneity: costs, quality, markups and product scope (*i.e.,* the number of products produced by firms). We develop a structural model of heterogeneous multiproduct firms that can be used to decompose the firm-size distribution into the relative contributions of each component. We first use this framework to structurally estimate elasticities of substitution across varieties between and within multiproduct firms. We next implement our model-based decomposition for the just over 50,000 firms that supply goods with barcodes in the Nielsen HomeScan Database in a typical quarter. This decomposition uses the structure of the model to isolate different margins in the data without making assumptions about how those margins are related to one another (as in the business cycle decomposition of Chari, Kehoe and McGratten 2007 in the macroeconomics literature). Our framework requires only price and expenditure data and hence is widely applicable. We separate out the contributions of price and quality using the exclusion restriction that marginal cost only affects firm sales through price. In contrast, quality is a demand shifter that shifts sales conditional on price.[1]

Our results point to quality differences, which we define as average consumer utility per physical unit of output, as being the principal reason why some firms are successful in the marketplace and others are not. Depending on the specification considered, we find that 50-70 percent of the variance in firm size can be attributed to differences in quality, about 23-30 percent to differences in product scope, and less than 24 percent to cost differences. When we turn to examine time-series evidence, the results become even more stark. Virtually all firm growth can be attributed quality improvements with most of the remainder due to increases in scope. These results suggest that most of what economists call differences in revenue productivity reflects differences in quality rather than cost.

Our framework uses a nested constant elasticity of substitution (CES) utility system that allows the elasticity of substitution between varieties within a firm to differ from the elasticity of substitution between varieties supplied by different firms. Our choice of this CES demand structure is guided by its prominence, tractability and empirical feasibility. Across international trade, economic geography, and macroeconomics, there is little doubt that this framework is the preferred approach to modeling product variety. Since our approach nests the standard CES-monopolistic competition model as a

[1]Our definition of quality is standard in the literature. For example Sutton (1986) defines "'vertical' product differentiation...has the defining property that if two distinct products are offered at the same price, then all consumers prefer the same one (the higher-quality product)"

special case, we can compare our results with those that would obtain in a standard trade, macro or economic geography model. However, we generalize this standard model to allow firms to supply multiple products (a pervasive feature of our data) and to have market power (since the largest firms in our data are far from being measure zero).

Incorporating these features into our structural model yields a number of additional insights. Our model makes clear a conceptual problem in the estimation of firm productivity that is likely to bias existing estimates. Most productivity estimates rely on the concept of real output, which is calculated by dividing nominal output by a price index. However, the formula for any economically motivated price index, which is the same as a unit expenditure function, is dependent on implicit assumptions about how the output of firms enters utility. *Thus, one cannot move from nominal output to real output without imposing assumptions about the structure of the demand system.*[2] Our results show theoretically and empirically that the CES measure of a multiproduct firm's price is highly sensitive to how differentiated its output is and how many products it supplies. The sensitivity of the CES price index to demand parameters, such as the elasticity of substitution and whether multiproduct firms exist, implies that estimates of real output are equally sensitive to these demand parameters. We show that if demand has a nested CES structure, conventional measures of real output will have a downward bias that rises with firm size with an elasticity of around one third. In other words, real output variation is substantially greater than nominal output variation. This bias also implies that true productivity differences are much larger than conventionally measured productivity differences.

The bias is driven by two features of reality that are typically ignored in most analyses. First, most analyses treat producers as single-product firms so they can avoid complications arising from the challenges of measuring the real output of multiproduct firms. However, our results indicate that multiproduct firms are the norm. For example, we document that 70 percent of firms supply more than one barcode and these firms account for more than 99 percent of output in their sectors.[3] Therefore truly single product firms account for a negligible share of sales in our data. Second, we show that if the output of multiproduct firms is differentiated, the common assumption that total firm output is simply the sum of the output of each good understates real output for multiproduct firms, and the degree of this downward bias rises in the number of products supplied.

Our framework also provides a new metric for quantifying the extent to which a firm's products are differentiated from those of its rivals. If a firm's products are perfectly substitutable with each other but not with those of other firms, then 100 percent of a new product's sales will come from the firm's existing sales, which implies a cannibalization rate of one. However, if a firm's market share is negligible (as it is for most firms) and its products are as differentiated from each other as they are from products supplied by other firms, none of a new product's sales will come at the

[2]This point was stressed by Aristotle in his *Nicomachean Ethics* (Book V, Section 5): "Demand holds things together as a single unit.... In truth it is impossible that things differing so much should become commensurate, but with reference to demand they may become so." (see http://classics.mit.edu/Aristotle/nicomachaen.5.v.html)

[3]By contrast, U.S. Census of Manufactures data indicates that producers of multiple five-digit Standard Industrial Classification (SIC) products account for 37 percent of firms and 87 percent of shipments (Bernard, Redding and Schott 2010). The difference comes from defining single product firms as one-industry firms as opposed to one-barcode firms.

expense of the firm's other products, which implies a cannibalization rate of zero. We estimate that the cannibalization rate for the typical firm is about 50 percent, indicating that although products supplied by the same firm are more substitutable with each other than with those of other firms, it is not correct to think of them as perfect substitutes.

We find that the monopolistic competition benchmark of atomistic firms with constant markups provides a good approximation for the vast majority of firms. The reason is simple. Most firms have trivial market shares and hence are unable to exploit their market power. However, there is substantial variation in markups for the very largest firms that account for disproportionate shares of aggregate sales. This variation is greater under quantity competition than under price competition. In most sectors, the largest firm has a market share above 20 percent, which enables it to charge a markup that is thirty percent higher than that of the median firm under price competition and double that of the median firm under quantity competition. We use the model to undertake counterfactuals, in which we show that these variable markups for the largest firms have quantitatively relevant effects on aggregate welfare and the firm sales distribution.

The remainder of the paper is structured as follows. Section 2 reviews the related literature. Section 3 discusses the data. Section 4 introduces the model. Section 5 uses the structure of the model to derive moment conditions to estimate elasticities of substitution and undertake our decomposition of firm sales. Section 6 presents our estimation results. Section 7 undertakes counterfactuals. Section 8 concludes.

2 Related Literature

Over the last decade, the fields of international trade and macroeconomics have undergone a transformation as the dissemination of micro datasets and the development of new theories has led to a shift in attention towards firm heterogeneity. Existing research has suggested a number of candidate explanations for differences in firm performance, including differences in production efficiency (e.g. Melitz 2003), product quality (e.g. Eslava, Fieler and Xu 2014, Johnson 2012, Khandelwal 2010, Schott 2004), markups (e.g. De Loecker and Warzynski 2012, De Loecker, Goldberg, Khandelwal and Pavcnik 2014), fixed costs (e.g. Das, Roberts and Tybout 2009), and the ability to supply multiple products (e.g. Arkolakis and Muendler 2010, Bernard, Redding and Schott 2010, 2011, Eckel, Iacavone, Javorcik and Neary 2013, Eckel and Neary 2010, Mayer, Melitz and Ottaviano 2014). While existing research typically focuses on one or more of these candidate explanations, they are all likely to operate to some degree in the data. We develop a general theoretical model that incorporates each of these candidate explanations and estimate that model structurally using disaggregated data on prices and sales by firm and product. We use the estimated model to provide evidence on the quantitative importance of each source of firm heterogeneity and the ways in which they interact with one another.

In much of the literature on firm heterogeneity following Melitz (2003) productivity and product quality are isomorphic. Under the assumption of CES preferences and monopolistic competition, productivity and product quality enter equilibrium firm revenue in exactly the same way. However, these different sources of firm heterogeneity have different implications for firm revenue conditional on prices (e.g. Berry 1994, Khandelwal 2010). While productivity only affects firm revenue through prices, product quality is a demand-shifter that shifts firm revenue conditional on prices. An advantage of our approach is that we observe prices and sales in our data, and hence we are able to separate these two sources of dispersion of firm sales.[4]

Most of the existing research on firm heterogeneity in trade and macroeconomics has assumed that firms are atomistic and compete under conditions of monopolistic competition (e.g. Melitz 2003 and Melitz and Ottaviano 2008). In contrast, a small number of papers have allowed firms to be large relative to the markets in which they operate (e.g. Atkeson and Burstein 2008, Amiti, Itskhoki and Konings 2012, and Edmond, Midrigan and Xu 2012). When firms internalize the effects of their decisions on market aggregates, they behave systematically differently from atomistic firms. Even under CES demand, firms charge variable mark-ups, because each firm internalizes the effects of its pricing decisions on market price indices and these effects are greater for larger firms.[5] Furthermore, since firms are of positive measure, idiosyncratic shocks to these "granular" firms can affect aggregate outcomes, as in Gabaix (2011) and di Giovanni and Levchenko (2012). In contrast to these papers, we structurally estimate a model of heterogeneous firms, and show how it can be used to recover the determinants of the dispersion of firm sales and undertake counterfactuals.

To the extent that such large firms supply multiple products, they take into account the effect of introducing new varieties on the sales of existing varieties. Most of the existing theoretical research on multiproduct firms in trade and macroeconomics has abstracted from these cannibalization effects by again assuming atomistic firms (e.g. Agur 2010, Allanson and Montagna 2005, Arkolakis and Muendler 2012, Bernard, Redding and Schott 2010, 2011, Mayer, Melitz and Ottaviano 2014 and Nocke and Yeaple 2006). Important exceptions that explore cannibalization effects theoretically are Feenstra and Ma (2008), Eckel and Neary (2010), and Dhingra (2013). In contrast to these theoretical studies, we develop a structural model that can be used to provide quantitative evidence on how important it is to introduce such cannibalization effects into models of firm behavior.

Most existing empirical research on multiproduct firms has measured products using production classification codes (e.g. around 1,500 five-digit Standard Industrial Classification (SIC) categories in Bernard, Redding and Schott 2010) or trade classification codes (e.g. around 10,000 Harmonized System codes in Bernard, Jensen and Schott 2009). In contrast, we measure products at a much finer level of resolution using what we term "barcodes"—either 12-digit Universal Product Codes (UPCs)

[4] Our CES formulation of market demand can be derived from a discrete choice model of the demands of individual consumers, as shown in Anderson, de Palma and Thisse (1992).

[5] Our model generates variable mark-ups without a "choke price" above which demand is zero. Therefore this model lies outside the classes considered by Arkolakis, Costinot and Rodriguez-Clare (2012) and Arkolakis, Costinot, Donaldson and Rodriguez-Clare (2012), in which aggregate statistics such as the trade share and trade elasticity are sufficient statistics for the welfare gains from trade.

or 13-digit European Article Numbers (EANs)—in scanner data.[6] This measure corresponds closely to the level at which product choice decisions are made by firms, because it is rare for an observable change in product attributes to occur without the introduction of a new barcode.

Our econometric approach builds on the literature estimating elasticities of substitution and quantifying the contribution of new varieties to welfare following Feenstra (1994) and Broda and Weinstein (2006). We extend this estimation approach to allow firms to be of positive measure relative to the market and to supply multiple products. We show how this extended approach can be used to recover demand heterogeneity, marginal cost heterogeneity, variable markups, and cannibalization effects. While a number of other studies have used scanner data, including Broda and Weinstein (2010) and Chevalier et al. (2003), so far these data have not been used to estimate a structural model of heterogeneous multiproduct firms and quantify the sources of dispersion in the firm-size distribution.

3 Data

Our data source is the Nielsen HomeScan database which enables us to observe price and sales information for millions of products with a barcode.[7] Barcode data have a number of advantages for the purpose of our analysis. First, since barcodes are inexpensive but provide sellers access to stores with scanners as well as internet sales, producers have a strong incentive to purchase barcodes for all products that have more than a trivial amount of sales.[8] This feature of the data means that it is likely we observe all products supplied by firms. Second, since assigning more than one product to a single barcode can interfere with a store's inventory system and pricing policy, firms have a strong incentive not to reuse barcodes. This second feature of the data ensures that in general identical goods do not have different barcodes. Thus, a barcode is the closest thing we have empirically to the theoretical concept of a good. Finally, since the cutoff size for a firm is to make a sale rather than an arbitrary number of workers, we actually observe something close to the full distribution of firms.

Nielsen collects its barcode data by providing handheld scanners to on average 55,000 households per year to scan each good purchased that has a barcode.[9] Prices are either downloaded from the store in which the good was purchased, or they are hand entered, and the household records any deals used that may affect the price. These households represent a demographically balanced sample of households in 42 cities in the United States. Overall, the database covers around 30 percent of all

[6]Recently, the 12-digit UPCs have been upgraded to 13-digit EAN-13s (European Article Numbers). The extra digit of the EAN-13 allows for more products and firms.

[7]Our results are calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business. Information on availability and access to the data is available at http://research.chicagobooth.edu/nielsen

[8]GS1 provides a company with up to 10 barcodes for a $250 initial membership fee and a $50 annual fee. There are deep discounts in the per barcode cost for firms purchasing larger numbers of them (see http://www.gs1us.org/get-started/im-new-to-gs1-us)

[9]The data for 2004 through 2006 come from a sample of 40,000 households, while the data for 2007 through 2011 come from a sample of 60,000 households.

expenditure on goods in the CPI.[10] We collapse the household dimension in the data and collapse the weekly purchase frequency in order to construct a national quarterly database by barcode on the total value sold, total quantity sold, and average unit value.[11]

Defining products as goods with barcodes has a number of advantages over defining goods by industry classifications. While industry classifications aggregate products produced by a firm within an industry, our data reveals that large firms typically sell hundreds of different products within even a narrowly-defined sector. In other words, while prior work equates multiproduct firms with multi-industry firms, our data hews extremely closely to what an economist would call a multiproduct firm. In principle, it could be appropriate to aggregate all output of a firm within an industry into a single good (Is whole milk the same as skim milk? Is a six-pack of soda the same as a two-liter bottle?) However, as we will show theoretically, the validity of this procedure depends on the elasticity of substitution between the products supplied by a firm, which is an empirical question.

Instead of relying on product data for a single industry, we observe virtually the entire universe of goods purchased by households in the sectors that we examine. Our database covers approximately 1.4 million goods purchased at some point by households in our sample. The data were weighted by Nielsen to correct for sampling error. For example, if the response rate for a particular demographic category is low relative to the census, Nielsen re-weights the averages so that the price paid and the quantity purchased is representative of the United States as a whole.

Nielsen organizes the barcodes into product groups according to where they would likely be stocked in a store. The five largest of our 100 product groups are carbonated beverages, pet food, paper products, bread and baked goods, and tobacco. Output units are common within a product group: typically volume, weight, area, length, or counts. Importantly, we deflate by the number of units in the barcode, so prices are expressed in price per unit (*e.g.*, price per ounce). When the units are in counts, we also deflate by the number of goods in a multipack, so, for instance, we would measure price per battery for batteries sold in multipacks. While about two thirds of these barcoded items correspond to food items, the data also contains significant amounts on information about non-food items like medications, housewares, detergents and electronics. The first few digits of the barcode identify the manufacturer.[12]

Table 1 presents descriptive statistics for our sample of firms and barcodes in the 100 product groups. We weight the data by the sales of the product group in each year and average across years

[10] For further discussion of the Nielsen data, see Broda and Weinstein (2010).

[11] One question that naturally arises is whether it is appropriate to think of US firms competing in a national market or in different regional markets. Handbury and Weinstein (forthcoming) argue there is not that much heterogeneity in the price levels of barcoded goods across cities. Similarly, we find that national and regional expenditure shares and prices are strongly correlated in the data. National brands and firms (brands and firms that sell in 10 or more cities) account for 97 and 98 percent of expenditure respectively in cities with more than 1,000 households in the AC Nielsen data. Regressing city prices in these cities on national prices, quarter-year dummies, and barcode fixed effects, we find a partial R^2 (*i.e.*, an R^2 not inclusive of the fixed effects) of 0.88 and 0.73 for price levels and price growth respectively. These results suggest that most of the variation in U.S. consumer prices reflects national shocks not local ones.

[12] Firm prefixes are usually seven digits but they can be as long as eleven digits, so we use GS1 data to map the barcodes into firm identifiers. We could not obtain a firm identifier for just under 5 percent of the barcodes, and so we dropped these products.

because sectors like carbonated beverages are much larger economically than sectors like "feminine hygiene." There are on average 518 firms in each product group with 90 percent of the product groups having more than 200 firms. We see enormous range in firm product scope. The median number of products supplied by a firm is 3 and the average is 13. On average, 670,000 different UPCs were sold each quarter.

One of the most striking facts displayed in this table is the degree of firm heterogeneity. This is manifest in the skewness of the size and barcode distributions. The largest firm in an industry typically sells 2600 times more than the median firm. We see similar patterns in terms of product scope and sales per product. The firm with the most products typically has 134 times more products than the firm with the median number of barcodes, and the barcode with the most sales on average generates almost almost a thousand times more revenue than the revenue of the median barcode.

Table 1: Sample Statistics

	Mean	Median	Std. Dev.	10th Percentile	90th Percentile	Max Value
No. Firms per Product Group	518	416	309	200	942	1434
Firm Sales	4141	152	25836	5	4599	402350
Log Firm Sales	12	12	3	8	15	19
No. of UPCs per Firm	13	3	32	1	32	405
UPC Sales	314	40	1224	2	650	36590

Note: Weighted by product group. Firm and UPC sales in thousands. Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

We see in Table 2 that almost 90 percent of sales in a product group was produced by firms with sales in the top decile of sales. Table 3 provides a more detailed description of this firm heterogeneity by focusing on the ten largest firms in each product group (where we weight the averages by the sales of the product group). Table 3 reveals an almost fractal nature of firm sales. Around two-thirds of all of the sales of firms in the top decile is produced by the ten largest firms (which on average only account for 1.9 percent of firms in this decile). While on average half of all output in a product group is produced by just five firms, 98 percent of firms have market shares of less than 2 percent. Thus, the typical sector is characterized by a few large firms and a vast competitive fringe comprised of firms with trivial market shares. A second striking feature of the data is that even the largest firms are not close to being monopolists. The largest firm in a product group on average only has a market share of 23 percent. Finally, the data reveal that firms in the top decile of sales are all multiproduct firms, firms on average supply 67 different goods, and the largest firms supplying hundreds of goods.

The extent of multiproduct firms can be seen more clearly in Table 4, which shows the results of splitting the data by the number of UPCs supplied by a firm. While single product firms constitute about one third of all firms on average, these firms account for less than one percent of all output. In other words virtually all output is supplied by multiproduct firms. Moreover, the fact that over ninety percent of output is sold by firms selling eleven or more varieties and two-thirds of all output is supplied by firms selling more than fifty varieties, suggests that single product firms are more the

Table 2: Size Distribution by Decile

Ranked Decile	Decile Market Share	Mean Firm Market Share	Mean Log Firm Sales	Avg. Std. Dev. log UPC Sales	Mean No. UPCs per Firm	Median No. UPCs per Firm
1	87.82	2.65	16.4	1.8	67.1	44.1
2	7.03	0.22	14.5	1.6	25.5	20.5
3	2.67	0.08	13.5	1.6	14.4	11.6
4	1.25	0.04	12.7	1.5	8.9	7.1
5	0.63	0.02	12.0	1.4	5.8	4.5
6	0.32	0.01	11.4	1.3	3.9	3.1
7	0.17	0.00	10.7	1.2	2.8	2.2
8	0.08	0.00	9.9	1.1	2.0	1.5
9	0.03	0.00	8.9	0.9	1.5	1.1
10	0.01	0.00	7.3	0.6	1.2	1.0

Note: Largest decile is ranked first. Weighted by product group. UPC counts and mean firm sales are across firm within each decile. Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

Table 3: Size Distribution by Firm Rank

Firm Rank	Firm Market Share (%)	Log Firm Sales	No. UPCs per Firm
1	22.7	19.4	296.7
2	12.5	18.8	193.8
3	7.8	18.4	155.2
4	5.4	18.1	119.5
5	4.1	17.8	115.5
6	3.3	17.6	111.2
7	2.8	17.4	105.3
8	2.4	17.2	92.8
9	2.1	17.1	85.8
10	1.9	17.0	73.3

Note: Largest firm is ranked first. Weighted by product group. Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

exception than the rule.

Large firms not only sell more products but they also sell a lot more of each product. The penultimate column of Table 4 documents that while the typical barcode sold by a single-product firm only only brings in $65,4000 in revenue, the typical barcode sold by a firm selling over one hundred barcodes brings almost twice as much ($122,500). In other words, large firms not only supply more products but they sell more of each product. If firms differed only in the fixed cost of adding new varieties, one would not expect to see large firms sell more of each variety. The fact that they do strongly suggests that large firms must also differ in the marginal cost or quality of their output.

Table 4: Size Distribution by Number of UPCs

No. of UPCs	No. of Firms	Share of Value (%)	Mean Sales (Thousands)	Median Sales (Thousands)	Avg. St. Dev. UPC Sales
1	165	0.9	73	65447	
2–5	171	3.7	327	69802	1.31
6–10	63	4.9	1207	56096	1.50
11–20	48	8.9	2924	66186	1.57
21–50	44	19.1	7823	66474	1.65
51–100	18	18.7	23385	119190	1.73
> 100	10	43.9	123156	122528	1.80

Note: Weighted by product group. To get last column, calculate standard deviation over log UPC sales by product group, then take weighted average across product groups. Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

In sum, our overview of the data reveals some key features that we model in our empirical exercise. First, the vast majority of firms have trivial market shares, which means that if we believe that all firms may have some market power, we need to work with demand systems that do not imply that firms with trivial market shares have trivial markups. Second, the fact that most economic output is produced by multiproduct firms impels us to build this feature directly into the estimation system and allow for both differences in the fixed cost of developing new products as well as cost and quality differences among products. Finally, the fact that there are firms with non-trivial market shares implies that at least some firms are likely to have market power. To the extent that these large firms internalize the effects of their price choices on market aggregates, this concentration of market shares will induce departures from the monopolistic competition benchmark.

4 Theoretical Framework

Our choice of functional forms is motivated not only by the data issues we identified above but also by some theoretical concerns. Since much of the theoretical literature in international trade and economic geography has worked with CES models, we want our results to nest this case (at least within product groups), so that our results easily can be compared with existing work in trade, macro and regional economics. However, we also need a framework that allows the elasticity of substitution for products supplied by the same firm to be different than that for products supplied by different firms, thereby to allow for the possibility of cannibalization effects. Finally, we need a setup that can be applied to firm-level data without imposing implausible assumptions or empirical predictions. This last requirement rules out two common demand systems: linear demand and the symmetric translog. A linear demand system would be problematic in our setting because one needs to impose the assumption that income elasticities are equal to zero and the estimation of marginal costs derived from a linear demand system can often result in negative values. While the symmetric translog demand system improves on the linear demand system in this regard, it is a difficult system to implement at the firm level because it has the undesirable result that firms with negligible market shares have negligible markups.

Our estimation strategy therefore is based on an upper level Cobb-Douglas demand system across product groups with CES nests below it. The upper-level Cobb-Douglas assumption dramatically simplifies the estimation process because its implication of constant product group expenditure shares allows us to assume that firms supplying different classes of products, e.g. carbonated beverages and pet food, do not interact strategically. However, the nested CES structure within broadly defined "product groups" allows for strategic interactions among firms supplying similar products. Within this structure the real consumption of each product group is composed of the real consumption of each firm's output, which itself is made up of the consumption of each of the varieties supplied by the firm.

4.1 Demand

In order to implement this approach, we assume that utility, U_t, at time t is a Cobb-Douglas aggregate of real consumption of each product group, C_{gt}:

$$U_t = \prod_{g \in G} C_{gt}^{\varphi_{gt}}, \qquad \sum_{g \in G} \varphi_{gt} = 1,$$

where g denotes each product group, φ_{gt} is the share of expenditure on product group g at time t.[13] Within product groups, we assume two CES nests for firms (f) and UPCs (u). The first nest for firms enables us to connect with the existing literature on measuring firm productivity. The second nest for UPCs enables us to incorporate multi-product firms. Therefore the consumption indices for product groups and firms can be written as follows for tier of utility $j \in \{g, f\}$:

$$C_{jt} = \left[\sum_{k \in K_t} \left(\varphi_{kt} C_{kt} \right)^{\frac{\sigma_{Kg}-1}{\sigma_{Kg}}} \right]^{\frac{\sigma_{Kg}}{\sigma_{Kg}-1}}, \qquad \sigma_{Kg} > 1,\ \varphi_{kt} > 0, \tag{1}$$

where k is the tier of utility below j (*i.e.*, if j corresponds to firm f, k corresponds to UPCs u, and if j corresponds to a product group g, k corresponds to firms f); K_t is the set of varieties k; C_{kt} denotes consumption of variety k; φ_{kt} is the perceived quality of variety k; σ_{Kg} is the constant elasticity of substitution across varieties k for product group g. In other words, the real consumption in any

[13]While we allow the Cobb-Douglas parameters φ_{gt} to change over time, we find that product-group expenditure shares are relatively constant over time, which suggests that a Cobb-Douglas functional form with time-invariant parameters would provide a reasonable approximation to the data.

product group, g, is a function of the consumption of each firm's output, C_{ft}, weighted by the quality consumers assign to that firm's physical output at time t, φ_{ft}, and adjusted for the substitutability of the output of each firm, σ_{Fg}.[14] Similarly, we can write the sub-utility derived from the consumption of a firm's output, C_{ft}, as a function of the consumption of each UPC (i.e. barcode) supplied by that firm, C_{ut}, multiplied by the quality assigned to the physical output of that barcode, φ_{ut}, and adjusted by the substitutability between the various varieties supplied by the firm, σ_{Ug}.[15]

There are a few features of this specification that are worth noting. First, if the elasticity of substitution across varieties supplied by a firm, σ_{Ug}, is finite, then the real output of a multiproduct firm is not equal to the sum of the outputs of each product. For example, if the only reason firms differ in size is that larger firms supply more varieties than smaller firms, then assuming firm real output is the sum of the output of each variety will tend to understate the relative size of larger firms. This size bias is a topic that we will explore in much more detail later. Second, for much of what follows, we are focused on the sales decomposition within a given product group (g), so for notational simplicity we can suppress the g subscript on σ_{Fg} and σ_{Ug} until we need it again to pool results across sectors in Section 4.6. Third, we would expect (but do not impose) that the elasticity of substitution across varieties is larger within firms than across firms, *i.e.*, $\sigma_U \geq \sigma_F$. When the two elasticities are equal, our system will collapse to a standard CES at the product-group level, and if the inequality is strict, we will show that our setup features cannibalization effects. While our nested CES specification provides a parsimonious and natural approach to modeling multi-product firms, we return to consider the robustness of our results to alternative nesting structures in subsection 6.6 below.

Third, we allow firms to be large relative to product groups (and hence internalize their effects on the consumption and price index for the product group). But we assume that the number of product groups is sufficiently large that each firm remains small relative to the economy as a whole (and hence takes aggregate expenditure E_t and factor prices as given). Finally, since the utility function is homogeneous of degree one in quality it is impossible to have a firm-quality, φ_{ft}, that is independent of the quality of the varieties produced by that firm, φ_{ut}. We therefore need to choose a normalization. It will prove convenient to normalize the geometric means of the φ_{ut} for each firm and the φ_{ft} for each product group to equal one:

[14] A small number of firms operate across multiple product groups. We assume that consumers view each of these firm-product-groups as a separate firm and that pricing and UPC introduction decisions are made for each firm-product-group separately. The Cobb-Douglas assumption for the upper tier of utility implies that no firm would have an incentive to price strategically across firm-product-groups, because product group expenditure shares are fixed by parameters. From now onwards, we refer to firm-product-groups as firms. These assumptions are also consistent with product groups being quite distinct from one another (e.g. Carbonated Beverages versus Office Supplies) and the vast majority of firms being active in only one product group.

[15] Our definition of quality is the utility per common physical unit (e.g. utility per ounce of a firm's output). However, variation in quality could either be interpreted as a difference in utility per physical unit or as variation in the number of identical-quality unobservable sub-units within a physical unit. For example, it is isomorphic to say that Firm A produces products with twice the utility per ounce as Firm B and to say that 1/2 an ounce of Firm A's product generates as much utility as an ounce of Firm B's product. In the latter case, an ounce of Firm A's product would contain two "1/2 ounce sub-units" each of which has identical quality to Firm B's product. We think our interpretation of utility per physical unit is the most natural for our data.

$$\left(\prod_{u\in U_{ft}}\varphi_{ut}\right)^{\frac{1}{N_{ft}}}=\left(\prod_{f\in F_{gt}}\varphi_{ft}\right)^{\frac{1}{N_{gt}}}=1, \tag{2}$$

where U_{ft} is the set of varieties supplied by firm f at time t, and N_{ft} is the number of elements of this set; F_{gt} is the set of firms in product group g at time t, and N_{gt} is the number of elements of this set. Thus, firm average quality (φ_{ft}) corresponds to a demand-shifter that affects all products supplied by the firm *proportionately*, while product quality (φ_{ut}) is a demand-shifter that determines the *relative* sales of individual products within the firm.

We can gain some intuition for this framework by by using the product group "carbonated beverages" as an example. Aggregate utility depends on the expenditure share (given by φ_{gt}) and amount of consumption of goods in the carbonated-beverages product group, (given by C_{gt}). The utility derived from the consumption of carbonated beverages depends on the quality of Coke versus Perrier (φ_{ft}), the amounts of each firm's real output consumed (C_{ft}), and the degree of substitutability between Coke and Perrier (σ_F). Finally, the real amount of Coke or Perrier consumed (C_{ft}) depends on the number of different types of soda produced by each company (U_{ft}), the quality of each of these types of soda (φ_{ut}), the consumption of each variety of soda (C_{ut}), and how similar varieties of Coke (or Perrier) products are with other varieties offered by the same company (σ_U).

It turns out that it also will be useful to also work with the exact price index for consumption:

$$P_{jt}=\left[\sum_{k\in K}\left(\frac{P_{kt}}{\varphi_{kt}}\right)^{1-\sigma_k}\right]^{\frac{1}{1-\sigma_k}}. \tag{3}$$

Without loss of generality, we index varieties k in a given year t from the largest in terms of sales to the smallest, and we denote the variety with the largest sales in a given year by $\underline{k}$.

Using the properties of CES demand, the expenditure share of variety k within tier j (S_{kt}) is equal to the elasticity of the price index for tier j with respect to the price of variety k and is given by the following expression:

$$S_{kt}=\frac{\left(P_{kt}/\varphi_{kt}\right)^{1-\sigma_K}}{\sum_{k\in K_t}\left(P_{kt}/\varphi_{kt}\right)^{1-\sigma_K}}=\frac{dP_{jt}}{dP_{kt}}\frac{P_{kt}}{P_{jt}}. \tag{4}$$

Equation (4) makes clear exactly how we conceive of quality in this setup. Holding fixed prices, a UPC with higher quality will have higher market shares. Similarly, holding fixed prices, firms that supply higher quality goods will have greater market shares.

The role of both firm and product quality also can be seen by writing down the demand for the output of each variety:

$$C_{ut}=\varphi_{ft}^{\sigma_F-1}\varphi_{ut}^{\sigma_U-1}E_{gt}P_{gt}^{\sigma_F-1}P_{ft}^{\sigma_U-\sigma_F}P_{ut}^{-\sigma_U}, \tag{5}$$

where E_{gt} denotes total expenditure on product group g at time t. Equation (5) is critical in determining how we can use this framework for understanding the different roles played by costs and quality for understanding the sales of a firm. While quality has a direct effect on consumer demand independent of price, cost only affects consumer demand through price. Thus, the specification of how cost and quality affect firm pricing decisions is crucial for our identification strategy.

4.2 Technology

We allow the costs of supplying products to the market to vary across UPCs, firm-product-groups and firms. This specification encompasses both heterogeneity in productivity across firms (as in Melitz 2003) and heterogeneity in productivity within firms (as in Bernard, Redding and Schott 2011). All costs are incurred in terms of a composite factor input that is chosen as our numeraire. We assume that the variable cost function is separable across UPCs and that supplying Y_{ut} units of output of UPC u incurs a total variable cost of $A_{ut}\left(Y_{ut}\right) = a_{ut}Y_{ut}^{1+\delta_g}$, where a_{ut} is a cost shifter and $\delta_g > 0$ parameterizes the convexity of marginal costs with respect to output. Once again, we will suppress the g subscript on δ_g until Section 4.6. In addition, each firm faces a fixed market entry cost of $H_{ft} > 0$ (e.g. the fixed costs of headquarters operations) and a fixed market entry cost for each UPC supplied of $h_{ft} > 0$ (e.g. the fixed costs of product development and distribution).

4.3 Profit Maximization

In our baseline specification, we assume that firms choose prices under Bertrand competition, though we also report results in which firms instead choose quantities under Cournot competition. Under our assumption of CES preferences, the decisions of any one firm only affect the decisions of other firms through the product group price indices (P_{gt}). Each firm chooses the set of UPCs $u \in \left\{\underline{u}_{ft}, \ldots, \overline{u}_{ft}\right\}$ to supply and their prices $\{P_u\}$ to maximize its profits:

$$\max_{N_{ft},\{P_{ut}\}} \Pi_{ft} = \sum_{u=\underline{u}_{ft}}^{\overline{u}_{ft}} \left[P_{ut}Y_{ut} - A_u\left(Y_{ut}\right)\right] - N_{ft}h_{ft} - H_{ft}, \tag{6}$$

where we index the UPCs supplied by the firm from the largest to the smallest in sales, and the total number of goods supplied by the firm is denoted by N_{ft}, where $\overline{u}_{ft} = \underline{u}_{ft} + N_{ft}$.

Multiproduct firms that are large relative to the market internalize the effect of their decisions for any one variety on the sales of their other varieties. From the first-order conditions for profit maximization, we can derive the firm markup for each UPC, as shown in Appendix A.1:

$$\mu_{ft} = \frac{\varepsilon_{ft}}{\varepsilon_{ft} - 1}, \tag{7}$$

where we define the firm's perceived elasticity of demand as

$$\varepsilon_{ft} = \sigma_F - \left(\sigma_F - 1\right) S_{ft} = \sigma_F \left(1 - S_{ft}\right) + S_{ft}, \tag{8}$$

and the firm's pricing rule as

$$P_{ut} = \mu_{ft}\gamma_{ut}, \qquad \gamma_{ut} = \left(1 + \delta\right) a_{ut} Y_{ut}^{\delta}, \tag{9}$$

where γ_{ut} denotes marginal cost.

One of the surprising features of this setup is that *markups only vary at the firm level within product groups*.[16] The intuition is that the firm internalizes that it is the monopoly supplier of its real output, which in our model equals real consumption of the firm's bundle of goods, C_{ft}. Hence its profit maximization problem can be thought of in two stages. First, the firm chooses the price index (P_{ft}) to maximize the profits from supplying real consumption (C_{ft}), which implies a markup at the firm level over the cost of supplying real output. Second, the firm chooses the price of each UPC to minimize the cost of supplying real output (C_{ft}), which requires setting the relative prices of UPCs equal to their relative marginal costs. Together these two results ensure the same markup across all UPCs supplied by the firm within a product group. However, for firms that operate across multiple product groups, prices are chosen separately for each product group. Therefore markups vary within firms across product groups.

The firm's perceived elasticity of demand (8) is less than the consumer's elasticity of substitution between firms, σ_F. The reason is that each firm is large relative to the market and hence internalizes the effect of its pricing choices on market price indices. When the firm raises the price of a UPC (P_{ut}) and hence the firm price index (P_{ft}), it reduces demand for that UPC and raises demand for the firm's other UPCs in (5). Furthermore, the rise in the firm price index also increases the product-group price index (P_{gt}), which raises demand for each UPC and for the firm as a whole in (5). Finally, our assumption of a Cobb-Douglas functional form for the upper tier of utility together with the assumption that each firm is small relative to the aggregate economy ensures that total expenditure on each product group (E_{gt}) in demand (5) is constant and unaffected by firm decisions.

Although consumers have constant elasticity of substitution preferences (σ_F), each firm perceives a variable elasticity of demand (ε_{ft}) that is decreasing in its expenditure share (S_{ft}), as in Atkeson and Burstein (2008) and Edmond, Midrigan and Xu (2012). As a result, the firm's equilibrium pricing rule (9) involves a variable markup (μ_{ft}) that is increasing in its expenditure share (S_{ft}). For a positive equilibrium price (9), we require that the perceived elasticity of demand (ε_{ft}) is greater than one (firms produce substitutes), which requires that the elasticity of substitution between firms (σ_F) is sufficiently large. As a firm's sales becomes small relative to the product group ($S_{ft} \to 0$), the mark-up (7) collapses to the standard constant mark-up of price over marginal cost under monopolistic competition with atomistic multiproduct firms and nested CES preferences (see for example Allanson and Montagna 2005 and Arkolakis and Muendler 2010).

Using the above equilibrium pricing rule, overall UPC profits (Π_{ut}) are equal to UPC variable profits ($\pi_{ut}\left(N_{ft}\right)$) minus fixed costs. UPC variable profits in turn can be written in terms of UPC revenues ($P_{ut}Y_{ut}$), the markup (μ_{ft}), and the elasticity of costs with respect to output (ζ_u):

$$\Pi_{ut} = \pi_{ut}\left(N_{ft}\right) - h_{ft}, \tag{10}$$

$$\pi_{ut}\left(N_{ft}\right) = P_{ut}Y_{ut} - A_u\left(Y_{ut}\right) = \left(\frac{\zeta_u \mu_{ft} - 1}{\zeta_u \mu_{ft}}\right) P_{ut}Y_{ut}, \text{ where } \zeta_u = \frac{dA_u\left(Y_{ut}\right)}{dY_{ut}} \frac{Y_{ut}}{A_u\left(Y_{ut}\right)} = 1 + \delta. \tag{11}$$

That is, $\pi_{ut}\left(N_{ft}\right)$ denotes the variable profits from UPC u when the firm supplies N_{ft} UPCs.

[16]This result would also hold if we substituted nested logit demand for nested CES.

4.4 Cannibalization Effects

The number of UPCs supplied by each firm, N_{ft}, is determined by the requirement that the increase in profits from introducing an additional UPC, $\overline{u}_{ft}+1$, minus the reduction in profits from reduced sales of existing UPCs $u \in \left\{\underline{u}_{ft}, \ldots, \overline{u}_{ft}\right\}$ is less than the fixed cost of introducing the new UPC, h_{ft}. If a firm makes N_{ft} products in equilibrium, then it must be the case that if it were to introduce a new good, its profits would fall, *i.e.*,

$$
\begin{aligned}
\sum_{u=\underline{u}_{ft}}^{\overline{u}_{ft}+1} \pi_{ut}\left(N_{ft}+1\right)-\left(N_{ft}+1\right) h_{ft} \quad &< \quad \sum_{u=\underline{u}_{ft}}^{\overline{u}_{ft}} \pi_{ut}\left(N_{ft}\right)-N_{ft} h_{ft} \\
\Leftrightarrow \pi_{\overline{u}_{ft}+1,t}\left(N_{ft}+1\right)-\sum_{u=\underline{u}_{ft}}^{\overline{u}_{ft}}\left\{\pi_{ut}\left(N_{ft}\right)-\pi_{ut}\left(N_{ft}+1\right)\right\} \quad &< \quad h_{ft}.
\end{aligned} \tag{12}
$$

In the case where the number of UPCs is large and can be approximated by a continuous variable, we obtain after some manipulation (see Appendix A.2) an expression for the "cannibalization rate":

$$
-\frac{\partial Y_{ut}}{\partial N_{ft}} \frac{N_{ft}}{Y_{ut}} \quad = \quad \left[\left(\frac{\sigma_U-\sigma_F}{\sigma_U-1}\right)+\left(\frac{\sigma_F-1}{\sigma_U-1}\right) S_{ft}\right] S_{N_{ft}} N_{ft}>0, \ for \, \sigma_u \geq \sigma_f>1. \tag{13}
$$

where $S_{N_{ft}}$ is the share of firm revenues from the new product when firm supplies N_{ft} products, and S_{ft} is the share of consumer expenditure on firm f. This cannibalization rate is defined as the partial elasticity of the sales of existing products with respect to the number of products. This partial elasticity captures the direct effect of the introduction of a new product on the sales of existing products, through the firm and product group price indices, holding constant the prices and marginal costs of these existing products.

The first term on the right hand-side captures the cannibalization rate within firms: The introduction of a UPC reduces the firm price index (P_{ft}), which reduces the revenue of existing UPCs if varieties are more substitutable within firms than across firms ($\sigma_U > \sigma_F$). The second term captures the cannibalization rate across firms: The introduction of the new UPC reduces the product-group price index (P_{gt}), which reduces the revenue of existing UPCs if varieties are more substitutable within product-groups than across product-groups ($\sigma_F > 1$).

There are two useful benchmarks for understanding the magnitude of the cannibalization rate in equation (13). In both cases, it useful to think of the introduction of a "standardized" product that has a market share equal to the average market share of the firms other goods ($S_{N_{ft}}N_{ft} = 1$). If firms are monopolistic competitors, $S_{ft} \approx 0$. Moreover, if we also assume that all products supplied by a firm are as differentiated among themselves as they are with the output of other firms, *i.e.*, $\sigma_U = \sigma_F$, then the cannibalization rate will be zero because all sales revenue arising from introducing a new product will come from the sales of goods supplied by other firms. Thus, a world with monopolistic competition and equal product differentiation is a world with no cannibalization. Clearly, the cannibalization rate will rise if firms cease being small $S_{ft} > 0$ or if goods supplied by the same firm are more substitutable with each other than with goods supplied by different firms: $\sigma_U > \sigma_F$. At the other extreme, we can assume that goods are "perfect substitutes" within firms, *i.e.*, $\sigma_U = \infty > \sigma_f$, so that varieties are differentiated across firms but there is no difference between varieties supplied by the same firm. In this case the cannibalization rate will be 1 because any sales of a new product will be exactly offset by a reduction in the sales of existing products. Thus the cannibalization rate provides a measure of where in the spectrum ranging from perfect substitutes to equal differentiation the output of a firm lies.

4.5 The Sources of Firm Heterogeneity

In this section, we use the model to quantify the contribution of the different sources of firm heterogeneity to the dispersion in sales across firms. Nominal firm sales, E_{ft}, is the sum of sales across UPCs supplied by the firm:

$$E_{ft} \equiv \sum_{u \epsilon U_{ft}} P_{ut} C_{ut}.$$

Using CES demand (5), firm sales can be re-written as:

$$E_{ft} = \varphi_{ft}^{\sigma_F - 1} E_{gt} P_{gt}^{\sigma_F - 1} P_{ft}^{\sigma_U - \sigma_F} \sum_{u \epsilon U_{ft}} (\frac{P_{ut}}{\varphi_{ut}})^{1-\sigma_U}. \tag{14}$$

Using the firm price index (3) to substitute for $P_f^{\sigma_U - \sigma_F}$ and taking logarithms, we obtain:

$$\ln E_{ft} = (\sigma_F - 1) \ln \varphi_{ft} + \ln E_{gt} + (\sigma_F - 1) \ln P_{gt} + \left(\frac{1 - \sigma_F}{1 - \sigma_U} \right) \ln \left(\sum_{u \epsilon U_{ft}} \left(\frac{P_{ut}}{\varphi_{ut}} \right)^{1-\sigma_U} \right),$$

where the final term summarizes the impact of productivity, markups, UPC quality, and the number of products on firm sales. This final term can be further decomposed into the contributions of the number of UPCs supplied by a firm (N_{ft}), the geometric mean of marginal costs ($\tilde{\gamma}_{ft}$), relative quality-adjusted marginal costs across UPCs ($\left(\gamma_{ut} / \tilde{\gamma}_{ut}\right) / \varphi_{ut}$), and the firm markup ($\mu_{ft}$):

$$\begin{aligned} \ln E_{ft} = & \left\{ \ln E_{gt} + (\sigma_F - 1) \ln P_{gt} \right\} \\ & + \left\{ (\sigma_F - 1) \ln \varphi_{ft} + \left(\frac{\sigma_F - 1}{\sigma_U - 1} \right) \ln N_{ft} \right\} \\ & + \left\{ \left[- (\sigma_F - 1) \ln \tilde{\gamma}_{ft} \right] + \left(\frac{\sigma_F - 1}{\sigma_U - 1} \right) \ln \left(\frac{1}{N_{ft}} \sum_{u \epsilon U_{ft}} \left(\frac{\gamma_{ut} / \tilde{\gamma}_{ft}}{\varphi_{ut}} \right)^{1-\sigma_U} \right) \right\} + (1 - \sigma_F) \ln \mu_{ft}, \end{aligned} \tag{15}$$

where

$$\tilde{\gamma}_{ft} = \left(\prod_{u \in U_{ft}} \gamma_{ut} \right)^{\frac{1}{N_{ft}}}$$

and recall that $\gamma_{ut} = (1 + \delta) \, a_{ut} Y_{ut}^{\delta}$ denotes marginal cost.

Equation (15) decomposes firm sales into seven terms that capture the various margins through which firms can differ in sales. Clearly these margins are related to one another, since both the number of products and the markup are endogenous to firm quality, product quality and marginal cost. Nonetheless, the decomposition (15) isolates the direct effects of firm quality, product quality and marginal cost from their indirect effects through the number of products and the markup.

Although our decomposition is exact for any co-movement in variables, it is easiest to obtain intuition for this equation if we consider one-at-a-time, small movements in each variable so that we can safely ignore interactions between different variables. We therefore will explain the intuition for this equation in terms of small movements in each variable, and note that in general and in our empirical implementation we will allow all variables to move simultaneously.

The seven terms in equation (15) can be grouped into four main elements. The first element, contained in the first set of braces, captures market size and relative pricing. Our demand system is homogeneous of degree one in product group expenditures, so firm sales rise one to one with aggregate expenditures. The second term captures the impact of the product group price index that summarizes the prices of competing varieties. Holding fixed a firm's characteristics, an increase in the product group price level of one percent will cause the firm's sales to rise by $(\sigma_F - 1)$ percent. Here, the elasticity of substitution between firm's output and the output of other firms, σ_F, plays the crucial role of explaining how much a relative price movement affects firm sales.

"Total firm quality" is captured by the second two terms in braces, which capture the impacts of the average quality of the firm's output, "firm quality,"and the number of products the firm offers holding fixed average quality, "product scope". Consider two firms that supply the same number of products, but one firm supplies higher quality varieties (measured in utility per unit), meaning that $\varphi_{ft} > \varphi_{f't}$ and $N_{ft} = N_{f't}$. Firm f will then have a higher market share; how much depends on the elasticity of substitution between firm output, σ_F. For a larger value of this elasticity, a given difference in firm average quality will translate into a larger difference in firm market share.

Now consider two firms that supply products of identical quality but one firm supplies more UPCs than another ($N_{ft} > N_{f't}$). Here, it is easiest to think about this term in a symmetric world in which all goods and firms have identical quality ($\varphi_{ut} = \varphi_{ft} = 1$) and identical marginal cost ($\gamma_{ut} = \gamma_t$), so we can just focus on the role played by product scope. Although firms have identical qualities, they do not have identical market shares because they differ in the number of products they offer, *i.e.*, $\ln N_{ft} > \ln N_{f't}$. For example, if consumers treated all UPCs identically regardless of which firm supplied them, *i.e.*, $\sigma_U = \sigma_F$, firm f would sell $\ln\left(N_{ft}/N_{f't}\right)$ percent more output than firm f'. More generally, if the products supplied by a firm are more substitutable with each other than with those of other firms, $\sigma_U > \sigma_F$, the percentage gain in sales accruing to a firm that adds a product will be less than one reflecting the fact that a new product will cannibalize the sales of its existing products. Indeed the degree of cannibalization will depend on the magnitude of σ_U; as this elasticity approaches infinity, the cannibalization rate will approach one, and all sales of new products will come from the sales of the firm's existing products. Hence, in this limiting case of $\sigma_u \to \infty$, adding product scope will have no impact on sales.

The terms in the third set of braces capture the role played by marginal costs. These costs can be divided into average marginal cost ($\tilde{\gamma}_{ft}$) and "cost dispersion". Average marginal costs ($\tilde{\gamma}_{ft}$)

is the more conventional measure which captures the fact that high cost firms have lower sales in equilibrium. The second term captures the fact that a firm sells more in equilibrium as the *dispersion* in the cost-to-quality ratio across its products increases. The intuition is straightforward. The term in logs is a form of Theil index of dispersion with the numerator being the cost of production of each variety relative to the average and the denominator being the quality of the variety relative to the average (which is always normalized to one). Consider a firm that supplies N_{ft} varieties of identical quality at identical cost so that $\varphi_{ut} = \gamma_{ut} = \widetilde{\gamma}_{ft} = 1$. Now consider a comparative static in which we multiply γ_{ut} by $\tau > 1$ and divide γ_{kt} for all $k \neq u$ by $\tau^{N_{ft}-1}$. This change has the property of keeping average marginal costs, $\widetilde{\gamma}_{ft}$, equal to one while increasing the dispersion in the cost of providing each element of the production bundle. Even though average costs are unaffected by construction, firm sales will rise because the increase in dispersion allows the firm to supply its production bundle more cheaply by shifting its output towards the sales of cheaper varieties. This cost-dispersion term indicates that a firm sells more if its costs relative to quality are less evenly distributed across its varieties. Here, we also see the first instance of the insidiousness of demand for understanding multiproduct firms—in the nested CES case, one cannot express firm-level marginal costs without reference to the demand parameters σ_F and σ_U.

Finally, the last term captures the role played by firm markups (μ_{ft}), which are themselves a function of the elasticity of substitution between firms (σ_F) and firm market share (S_{ft}).

4.6 Firm Sales Decompositions

While the decomposition given in equation (15) is exact for each sector, it is difficult to use that specification to understand the general determinants of firm size because aggregate product group expenditures are a major determinant of firm sales. Therefore we decompose firm sales in each product group relative to the geometric mean for that product group. Notationally, the algebra is clearer if we reintroduce the g subscript for all the relevant variables. We can write the sales decomposition for the average firm in product group g as

$$\begin{aligned}\overline{\ln E}_{gt} &= (\sigma_{Fg}-1)\overline{\ln\varphi}_{gt} - (\sigma_{Fg}-1)\overline{\ln\widetilde{\gamma}}_{gt} + \ln E_{gt} + (\sigma_{Fg}-1)\ln P_{gt} \\ &\quad + \left(\frac{1-\sigma_{Fg}}{1-\sigma_{Ug}}\right)\left[\overline{\ln N}_{gt} + (1-\sigma_{Ug})\,\overline{\ln\mu}_{gt} + \overline{\ln\left(\frac{1}{N_{uft}}\sum_{u\epsilon U_{ft}}\left(\frac{\gamma_{ut}/\tilde{\gamma}_{ft}}{\varphi_{ut}}\right)^{1-\sigma_{Ug}}\right)}_{gt}\right], \qquad (16)\end{aligned}$$

where we define $\bar{X}_{gt} = \frac{1}{F_t}\sum_{f\in g} X_{ft}$ and F_t is the number of firms in the product group. If we subtract equation (16) from equation (15) we obtain

$$
\begin{aligned}
\ln E_{ft} - \overline{\ln E}_{gt} \quad = \quad & (\sigma_{Fg}-1)\left\{\ln \varphi_{ft} - \overline{\ln \varphi}_{gt}\right\} + \left(\frac{1-\sigma_{Fg}}{1-\sigma_{Ug}}\right)\left[\ln N_{ft} - \overline{\ln N}_{gt}\right] \\
& -(\sigma_{Fg}-1)\left\{\ln \tilde{\gamma}_{ft} - \overline{\ln \tilde{\gamma}}_{gt}\right\} + (1-\sigma_{Fg})\left(\ln \mu_{ft} - \overline{\ln \mu}_{gt}\right) \\
& + \left(\frac{1-\sigma_{Fg}}{1-\sigma_{Ug}}\right)\left[\ln\left(\frac{1}{N_{uft}}\sum_{u\epsilon U_{ft}}\left(\frac{\gamma_{ut}/\tilde{\gamma}_{ft}}{\varphi_{ut}}\right)^{1-\sigma_{Ug}}\right) - \overline{\ln\left(\frac{1}{N_{uft}}\sum_{u\epsilon U_{ft}}\left(\frac{\gamma_{ut}/\tilde{\gamma}_{ft}}{\varphi_{ut}}\right)^{1-\sigma_{Ug}}\right)}_{gt}\right],
\end{aligned} \tag{17}
$$

which can be written more compactly as

$$
\begin{aligned}
\Delta^g \ln E_{ft} = & \left\{(\sigma_{Fg}-1)\Delta^g \ln \varphi_{ft} + \left(\frac{1-\sigma_{Fg}}{1-\sigma_{Ug}}\right)\Delta^g \ln N_{ft}\right\} \\
& + \left\{\left[-\left(\sigma_{Fg}-1\right)\Delta^g \ln \tilde{\gamma}_{ft}\right] + \left(\frac{1-\sigma_{Fg}}{1-\sigma_{Ug}}\right)\Delta^g \ln\left(\frac{1}{N_{ft}}\sum_{u\epsilon U_{ft}}\left(\frac{\gamma_{ut}/\tilde{\gamma}_{ft}}{\varphi_{ut}}\right)^{1-\sigma_U}\right)\right\} \\
& + \left(1-\sigma_{Fg}\right)\Delta^g \ln \mu_{ft}.
\end{aligned} \tag{18}
$$

The left-hand side of equation (18) is the sales of firm f relative to average firm sales in the sector, and the five terms on the right-hand side tell us the importance of relative firm average quality, scope, average marginal costs, cost dispersion, and markups in understanding cross-sectional differences in firm size. Moreover, we can also undertake this decomposition for a firm's sales growth rate by taking the first difference of equation (18), which in our notation simply involves replacing Δ^g with $\Delta^{g,t}$ in the equation. We now can decompose the cross-sectional variation in firm sales using a procedure analogous to Eaton, Kortum, and Kramarz's (2004) variance decomposition commonly used in the international trade literature. In particular, we regress each of the components of log firm sales on log firm sales as follows:

$$
\begin{aligned}
(\sigma_{Fg}-1)\Delta^g \ln \varphi_{ft} \quad &= \quad \alpha_\varphi \Delta^g \ln E_{ft} + \varepsilon_\varphi, && (19)\\
-\left(\sigma_{Fg}-1\right)\Delta^g \ln \tilde{\gamma}_{ft} \quad &= \quad \alpha_\gamma \Delta^g \ln E_{ft} + \varepsilon_\gamma, && (20)\\
\left(\frac{1-\sigma_{Fg}}{1-\sigma_{Ug}}\right)\Delta^g \ln\left[\frac{1}{N_{ft}}\sum_{u\epsilon U_{ft}}\left(\frac{\gamma_{ut}/\tilde{\gamma}_{ft}}{\varphi_{ut}}\right)^{1-\sigma_U}\right] \quad &= \quad \alpha_D \Delta^g \ln E_{ft} + \varepsilon_a, && (21)\\
\left(\frac{1-\sigma_{Fg}}{1-\sigma_{Ug}}\right)\Delta^g \ln N_{ft} \quad &= \quad \alpha_N \Delta^g \ln E_{ft} + \varepsilon_N, && (22)\\
\left(1-\sigma_{Fg}\right)\Delta^g \ln \mu_{ft} \quad &= \quad \alpha_\mu \Delta^g \ln E_{ft} + \varepsilon_\mu, && (23)
\end{aligned}
$$

where each firm's log sales is differenced relative to the largest firm in its product group. By the properties of OLS, $\alpha_\varphi + \alpha_\gamma + \alpha_D + \alpha_N + \alpha_\mu = 1$. The values for each of the α's provide us with a measure of how much of the variation in the distribution of firm sales can be attributed to each factor. If we replace Δ^g with $\Delta^{g,t}$ in the above equations, we can determine how much of the variance in sales growth rates can be attributed to each factor. Thus, estimation of these five equations provides us with a simple way to decompose firm sales in the cross section and in the time series.

4.7 Decomposing Changes in firm average quality

When decomposing firm growth, we can go one step further and understand how much of a firm's change in quality is due to the introduction of new products and how much is due to demand factors for existing products. In order to see how to do this decomposition, note that our normalization for UPC quality implies that the geometric mean of the φ_{ut} equals one for each firm in each period, *i.e.*,

$$\varphi_{ft} = \varphi_{ft} \left(\prod_{u \in U_{ft}} \varphi_{ut} \right)^{\frac{1}{N_{ft}}} \tag{24}$$

We can then can take log differences of this equation over time to yield:

$$\Delta^{g,t} \ln \varphi_{ft} \equiv \Delta \ln \varphi_{ft} - \Delta \overline{\ln \varphi}_{gt} = \Delta \ln \varphi_{ft} + \frac{1}{N_{ft}} \sum_{u \in U_{ft}} \ln \varphi_{ut} - \frac{1}{N_{f,t-1}} \sum_{u \in U_{f,t-1}} \ln \varphi_{u,t-1} - \Delta \overline{\ln \varphi}_{gt}, \tag{25}$$

where our normalization (2) implies that the second and third terms are both equal to zero.

To isolate the sources of changes in firm average quality, we distinguish between UPCs that are supplied in both periods versus those that are supplied in only one of the two periods. Let $I_{ft} = U_{ft} \cap U_{f,t-1}$ denote the set of UPC's that are supplied by firm f in both periods t and $t-1$. Similarly, define U_{ft}^{+} to be the set of newly introduced UPCs, *i.e.*, the set of UPCs in U_{ft} but not in $U_{f,t-1}$, and U_{ft}^{-} to be the set of disappearing UPCs, *i.e.*, the set of UPCs in $U_{f,t-1}$ but not in U_{ft}.

Noting that $\left|U_{ft}\right| = \left|I_{ft}\right| + \left|U_{ft}^{+}\right|$ and $\left|U_{ft-1}\right| = \left|I_{ft}\right| + \left|U_{ft}^{-}\right|$, we can decompose the change in firm average quality into quality changes for a constant set of products and quality changes from the adding and dropping of UPCs:

$$\begin{aligned} \Delta^{g,t} \ln \varphi_{ft} = & \left\{ \frac{\left|I_{ft}\right|}{\left|U_{ft}\right|} \left[\frac{1}{\left|I_{ft}\right|} \sum_{u \in I_{ft}} \ln \varphi_{ut} \right] - \frac{\left|I_{ft}\right|}{\left|U_{f,t-1}\right|} \left[\frac{1}{\left|I_{ft}\right|} \sum_{u \in I_{ft}} \ln \varphi_{u,t-1} \right] + \left(\Delta \ln \varphi_{ft} - \Delta \overline{\ln \varphi}_{gt} \right) \right\} \\ & + \left\{ \frac{\left|U_{ft}^{+}\right|}{\left|U_{ft}\right|} \left[\frac{1}{\left|U_{ft}^{+}\right|} \sum_{u \in U_{ft}^{+}} \ln \varphi_{ut} \right] - \frac{\left|U_{ft}^{-}\right|}{\left|U_{f,t-1}\right|} \left[\frac{1}{\left|U_{ft}^{-}\right|} \sum_{u \in U_{ft}^{-}} \ln \varphi_{u,t-1} \right] \right\}. \end{aligned} \tag{26}$$

The first term in braces captures quality changes for a constant set of products, and includes the change in firm average quality ($\Delta \ln \varphi_{ft}$), as well as the difference between average product quality for the common set of products (in square brackets) weighted by their relative importance in the set of products in each year. This difference adjusts changes in our measure of firm quality for changes in our normalization due to the entry and exit of new goods. We refer to the first term in braces in equation (26) as the "demand effect". It represents the change in consumers perceived quality of the UPCs present in both periods. If the firm does not add or eliminate any products, the summation terms will be zero since our normalization (2) means that average log product quality must be zero in all periods.

However, if the set of products changes, the summation terms will not necessarily be zero. Product turnover can influence sales either through product upgrading, which is measured in the second term in braces in equation (26), as well as through changes in the number of products, which enters separately into our decomposition of log firms sales (15) above. The "product upgrading effect" (the second term in braces in (26) captures quality changes arising from the adding and dropping of products. Product upgrading depends on average product quality for the entering and exiting products (in square brackets) weighted by their relative importance in the set of products in each year. For example, suppose a firm introduces a new high-quality product and retires a low-quality product leaving the number of products unchanged. In this case, the second term in braces will be positive reflecting the fact that the firm upgraded the average quality of its product mix.

5 Structural Estimation

Our structural estimation of the model has two components. First, given data on expenditure shares and prices $\{S_{ut}, S_{ft}, P_{ut}\}$ and known values of the elasticities of substitution $\{\sigma_U, \sigma_F\}$, we show how the model can be used to determine unique values of firm average quality (φ_{ft}), product quality (φ_{ut}), and marginal cost shocks (a_{ut}) up to our normalization of quality. These correspond to structural residuals of the model that are functions of the observed data and parameters and ensure that the model exactly replicates the observed data. We use these structural residuals to implement our decomposition of firm sales from Section 4.5 with our observed data on expenditure shares and prices.

Second, we estimate the elasticities of substitution $\{\sigma_U, \sigma_F\}$ using a generalization of Feenstra (1994) and Broda and Weinstein (2006) to allow allow firms to be large relative to the markets in which they operate (which introduces variable markups) and to incorporate multiproduct firms (so that firm pricing decisions are made jointly for all varieties). This estimation uses moment conditions in the double-differenced values of the structural residuals $\{\varphi_{ft}, \varphi_{ut}, a_{ut}\}$ and also has a recursive structure. In a first step, we estimate the elasticity of substitution across UPCs within firms for each product group $\{\sigma_U\}$. In a second step, we use these estimates for UPCs to estimate the elasticity of substitution across firms for each product group $\{\sigma_F\}$.

5.1 Structural Residuals

We begin by showing that there is a one-to-one mapping from the observed data on expenditure shares and prices $\{S_{ut}, S_{ft}, P_{ut}\}$ and the model's parameters $\{\sigma_U, \sigma_F, \delta\}$ to the unobserved structural residuals $\{\varphi_{ut}, \varphi_{ft}, a_{ut}\}$.

Given known values for the model's parameters $\{\sigma_U, \sigma_F, \delta\}$ and the observed UPC expenditure shares and prices, we can use the expression for the expenditure share given in equation (4) to determine UPC qualities $\{\varphi_{ut}\}$ up to our normalization that the geometric mean of UPC qualities is equal to one. These solutions for UPC qualities and observed UPC prices can be substituted into the

CES price index (3) to compute firm price indices $\{P_{ft}\}$. These solutions for firm price indices and observed firm expenditure shares can be combined with the CES expenditure share (4) to determine firm qualities $\{\varphi_{ft}\}$ up to our normalization that the geometric mean of firm qualities is equal to one. Furthermore observed firm expenditure shares and the CES markup (7) are sufficient to recover firm markups $\{\mu_{ft}\}$. These solutions for markups and observed UPC prices and expenditures can be substituted into the CES pricing rule (9) to determine the marginal cost shock (a_{ut}).

Finally, our solutions for markups and observed UPC expenditures can be combined with CES variable profits (11) to obtain upper upper bounds to the fixed costs of supplying UPCs (h_{ft}) and the fixed costs of operating a firm (H_{ft}). The upper bound for UPC fixed costs for each product group (h_{ft}) is defined by the requirement that variable profits for the least profitable UPC within a product group must be greater than this fixed cost. Similarly, the upper bound for firm fixed costs for each product group (H_{ft}) is defined by the requirement that variable profits for the least profitable firm must be greater than this fixed cost.

This mapping from the observed data on expenditure shares and prices $\{S_{ut}, S_{ft}, P_{ut}\}$ and the model's parameters $\{\sigma_U, \sigma_F, \delta\}$ to the unobserved structural residuals $\{\varphi_{ut}, \varphi_{ft}, a_{ut}\}$ does not impose assumptions about the functional forms of the distributions for the structural residuals or about their correlation with one another. When we estimate the model's parameters $\{\sigma_U, \sigma_F, \delta\}$ below, we impose some identifying assumptions on the double-differenced values of these structural results but not upon their levels. Therefore, having recovered these structural residuals, we can examine the functional form of their distributions and their correlation with one another.

5.2 UPC Moment Conditions

We now discuss our methodology for estimating the elasticities of substitution $\{\sigma_U, \sigma_F\}$ and the elasticity of marginal costs with respect to output (δ), which uses moment conditions in double-differenced values of the structural residuals $\{\varphi_{ft}, \varphi_{ut}, a_{ut}\}$. This estimation again has a recursive structure. In a first step, we estimate the elasticity of substitution across UPCs within firms for each product group $\{\sigma_U\}$ and the marginal cost elasticity (δ). In a second step, we use these estimates for UPCs to estimate the elasticity of substitution across firms for each product group $\{\sigma_F\}$.

In the first step, we double-difference log UPC expenditure shares (4) over time and relative to the largest UPC within each firm to obtain the following equation for relative UPC demand:

$$\triangle^{k,t} \ln S_{ut} = (1-\sigma_U)\triangle^{k,t} \ln p_{ut} + \omega_{ut}, \tag{27}$$

where$\triangle^{k,t}$ is the double-difference operator such that $\triangle^{k,t} \ln S_{ut} = \triangle \ln S_{ut} - \triangle \ln S_{kt}$; u is a UPC supplied by the firm; k corresponds to the largest UPC supplied by the same firm (as measured by the sum of expenditure across the two years); S_{ut} and p_{ut} are directly observed in our data; $\omega_{ut} = (1-\sigma_U)\left[\triangle \ln \varphi_{kt} - \triangle \ln \varphi_{ut}\right]$ is a stochastic error. Since we double difference the market shares of two UPCs supplied by the same firm, we eliminate all demand shocks that are common across a firm's UPCs, which leaves only demand shocks that affect the sales of one a firm's UPCs relative to another. For example, random demand shocks and the timing of holidays, weekends, and meteorological events might affect the success of certain products in a firm's lineup relative to others.

Double-differencing the UPC pricing rule (9) enables us to obtain an equation for relative UPC supply. Using the cost function ($A_u(Y_u) = a_{ut} Y_{ut}^{1+\delta}$), and the relationship between output and revenue ($Y_{ut} = S_{ut}/P_{ut}$), the UPC pricing rule given in equation (9) can be re-written as:

$$P_{ut} = \mu_{ft}\gamma_{ut} = \mu_{ft}^{\frac{1}{1+\delta}} (1+\delta)^{\frac{1}{1+\delta}} a_{ut}^{\frac{1}{1+\delta}} S_{ut}^{\frac{\delta}{1+\delta}}.$$

Taking logs and double-differencing, we obtain the following equation for relative UPC supply:

$$\triangle^{k,t} \ln p_{ut} = \frac{\delta}{1+\delta} \triangle^{k,t} \ln S_{ut} + \kappa_{ut}, \tag{28}$$

where the markup μ_f has differenced out because it is the same across UPCs within the firm; S_{ut} and p_{ut} are again directly observed in our data; $\kappa_{ut} = \frac{1}{1+\delta} \left[\triangle \ln a_{ut} - \triangle \ln a_{\underline{ut}}\right]$ is a stochastic error. Since the firm markup differences out and we observe prices, our estimation approach is the same under either price or quantity competition, and hence is robust across these different forms of competition. Once again the double differencing within a firm eliminates shocks that are common across a firm's products, and the fact that a barcode uniquely identifies a product means that changes in observable product attributes will be manifest in a change of barcode but not in a change in a_{ut}. Therefore the supply-side shocks κ_{ut} correspond to factors that affect one variety supplied by firm but not another. These might be problems in individual plants, exchange rate movements in the countries supplying those products, etc. We assume that these idiosyncratic shocks to marginal cost, κ_{ut}, are orthogonal to the idiosyncratic demand shocks, ω_{ut}.

Following Broda and Weinstein (2006), the orthogonality of idiosyncratic demand and supply shocks defines a set of moment conditions (one for each UPC):

$$G(\beta_g) = \mathbb{E}_{\mathbb{T}} \left[v_{ut}(\beta_g)\right] = 0, \tag{29}$$

where $\beta_g = \begin{pmatrix} \sigma_U \\ \delta \end{pmatrix}$ and $v_{ut} = \omega_{ut}\kappa_{ut}$. For each product group, we stack all the moment conditions to form the GMM objective function and obtain:

$$\hat{\beta}_g = \arg\min_{\beta_g} \left\{G^*(\beta_g)' W G^*(\beta_g)\right\} \qquad \forall g, \tag{30}$$

where $G^*(\beta_g)$ is the sample analog of $G(\beta_g)$ stacked over all UPCs in a product group and W is a positive definite weighting matrix. As in Broda and Weinstein (2010), we weight the data for each UPC by the number of raw buyers for that UPC to ensure that our objective function is more sensitive to UPCs purchased by larger numbers of consumers.

The moment condition (29) for each UPC involves the expectation of the product of the double-differenced demand and supply shocks: $v = \omega_{ut}\kappa_{ut}$. From relative demand (27) and relative supply (28), this expectation depends on the variance of prices, the variance of expenditure shares, the covariance of prices and expenditure shares, and parameters. Our identifying assumption that this expectation is equal to zero defines a rectangular hyperbola in (σ_U, δ) space for each UPC, along which a higher value of σ_U has to be offset by a lower value of δ in order for the expectation to be equal to zero (Leontief 1929). Therefore, this rectangular hyperbola places bounds on the demand and supply elasticities for each UPC, even in the absence of instruments for demand and supply. Furthermore, if the variances for the double-differenced demand and supply shocks are heteroskedastic across UPCs, the rectangular hyperbolas are different for each pair of UPCs, and their intersection can be used to separately identify the demand and supply elasticities (Feenstra 1994). Consistent with these identifying assumptions in Broda and Weinstein (2006, 2010) and Feenstra (1994), we find that the double-differenced demand and supply shocks are in general heteroskedastic.[17]

5.3 Firm Moment Conditions

We use our estimates of the UPC elasticities of substitution $\{\sigma_U\}$ from the first step to solve for UPC quality $\{\varphi_{ut}\}$ and compute the firm price indices $\{P_{ft}\}$ using equations (3) and (4). In our second step, we double difference log firm expenditure shares (4) over time and relative to the largest firm within each product-group, $\underline{f}$, to obtain the following equation for relative firm market share:

$$\Delta^{\underline{f},t} \ln S_{ft} = (1-\sigma_F)\, \Delta^{\underline{f},t} \ln P_{ft} + \omega_{ft}, \tag{31}$$

where the stochastic error is $\omega_{ft} \equiv -(\sigma_F - 1)\, \Delta^{\underline{f},t} \ln \varphi_{ft}$.

Estimating equation (31) using ordinary least squares would be problematic because changes in firm price indices could be correlated with changes in firm average quality: $\mathrm{Cov}\left(\Delta^{\underline{f},t} \ln P_{ft},\ \Delta^{\underline{f},t} \ln \varphi_{ft}\right) \neq 0$. To find a suitable instrument for changes in firm price indices, we use the structure of the model to write changes in firm price indices in terms of the underlying UPC characteristics of the firm. Using the CES expenditure shares (4), we can write relative UPC expenditures in terms of relative UPC prices and relative UPC qualities:

$$\frac{S_{ut}}{\tilde{S}_{ft}} = \left(\frac{P_{ut}/\varphi_{ut}}{\tilde{P}_{ft}/\tilde{\varphi}_{ft}}\right)^{1-\sigma_U}, \qquad u \in U_f, \tag{32}$$

where here we compare each UPC to the geometric mean of UPCs within the firm, which is denoted by a tilde so that $\tilde{S}_{ft} = \exp\left\{\frac{1}{N_{uft}} \sum_{u\in U_{ft}} \ln S_{ut}\right\}$ and $\tilde{P}_{ft} = \exp\left\{\frac{1}{N_{uft}} \sum_{u\in U_{ft}} \ln P_{ut}\right\}$. Using this expression for relative expenditure shares to substitute for product quality (φ_{ut}) in the CES price index (3), we can write the firm price index solely in terms of observed relative expenditures and the geometric mean of UPC prices:

$$\ln P_{ft} = \ln \tilde{P}_{ft} + \frac{1}{1-\sigma_U} \ln \left[\sum_{u\in U_{ft}} \frac{S_{ut}}{\tilde{S}_{ft}}\right], \tag{33}$$

where we have used our normalization that $\tilde{\varphi}_{ft} = 1$.

[17] In a White test for heteroskedasticity, we are able to reject the null hypothesis of homoskedasticity at conventional significance levels for 94 percent of product groups.

Equation (33) decomposes the firm price index into two terms. The first term is entirely conventional: the geometric mean of the prices of all goods supplied by the firm. When researchers approximate this firm price index using firm-level unit values or the prices of representative goods supplied by firms they are essentially capturing this component of the firm price index. The second term is novel and arises because the conventional price indexes are not appropriate for multiproduct firms. The log component of the second term is a variant of the Theil index of dispersion.[18] If the shares of all products are equal, the Theil index will equal $\ln N_{ft}$, which is increasing the number of products supplied by the firm, N_{ft}. Since the Theil index is multiplied by $(1-\sigma_U)^{-1} \leq 0$, the firm's price index falls as the number of goods supplied by the firm rises. This Theil index also increases as the dispersion of the market shares across UPCs within the firm rises.

One can obtain some intuition for this formula by comparing it to the conventional price indexes commonly used in economics. Firm price indexes are typically constructed by making at least one of two critical assumptions—firms only supply one product ($N_{ft} = 1$), or the goods supplied by firms are perfect substitutes (*i.e.*, $\sigma_U = \infty$). Either of these assumptions is sufficient to guarantee that the firm's price level equals its average price level. However, both of these assumptions are likely violated in reality. The average price of a firm's output prices overstates the price level for a multiproduct firms because consumers derive more utility per dollar spent on a firm's production if that production bundle contains more products. *Thus, the (geometric) average firm price is a theoretically rigorous way to measure the price level of firms that produce perfect substitutes, it overstates the prices of firms that produce differentiated goods, and this bias will tend to rise as the number of products supplied by the firm increases.*

The structure of the model implies that the dispersion of the shares of UPCs in firm expenditure (S_{ut}) only affects the shares of firms in product group expenditure (S_{ft}) through the firm price indices (P_{ft}). Double-differencing equation (33) for the log firm price index over time and relative to the largest firm within each product-group, we obtain:

$$\Delta^{\underline{f},t} \ln P_{ft} = \Delta^{\underline{f},t} \ln \widetilde{P}_{ft} + \frac{1}{1-\sigma_U} \Delta^{\underline{f},t} \ln \left[\sum_{u \in U_{ft}} \frac{S_{ut}}{\widetilde{S}_{ft}} \right], \tag{34}$$

where the model implies that the second term on the right-hand side containing the shares of UPCs in firm expenditure is a suitable instrument for the double-differenced firm price index in (31). We find that this instrument is powerful in the first-stage regression (34), with a first-stage F-statistic for the statistical significance of the excluded exogenous variable that is substantially above the recommended threshold of 10 from Stock, Wright and Yogo (2002).

[18]The standard Theil index uses shares relative to simple average shares, while ours expresses shares relative to the geometric mean.

6 Estimation Results

We present our results in several stages. First, we present our elasticity and quality estimates and show that they are reasonable. Second we use these estimates to examine cannibalization, and finally we present our results on the sources of firm heterogeneity.

6.1 Estimated Elasticities of Substitution

Because we estimate 100 σ_U's and σ_F's, it would needlessly clutter the paper to present all of them individually. Table 5 shows goods supplied by the same firm are imperfect substitutes. For UPCs, the estimated elasticity of substitution ranges from 4.7 at the 5th percentile to 17.6 at the 95th percentile with a median elasticity of 6.9. These numbers are large compared with trade elasticities reflecting the fact that products supplied by the same firm are closer substitutes than products supplied by different firms. The median elasticity implies that a one percent price cut causes the sales of that UPC to rise by 6.9 percent.[19]

Table 5: Distribution of 100 GMM Estimates

Ranked Percentile	σ_U	σ_F	$\sigma_U - \sigma_F$	δ
1	24.4	12.8	16.8	0.62
5	17.6	8.5	9.4	0.42
10	14.0	7.3	6.6	0.30
25	8.6	5.1	4.0	0.20
50	6.9	3.9	2.6	0.16
75	5.4	3.1	2.2	0.09
90	5.0	2.6	1.8	0.06
99	4.4	2.1	1.2	0.02

Note: Percentiles are decreasing: largest estimate is ranked first. Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

The fact that we estimate the different products supplied by a given firm to be imperfect substitutes for each other has profound implications for how we should understand firm pricing and productivity. All productivity estimates are based on a concept of real output, which, in index number theory, equals nominal output divided by the minimum expenditure necessary to generate a unit of utility. According to equation (1) the quality-adjusted flow of consumption from a firm's output is $\varphi_{ft}C_{ft}$, so the corresponding expenditure function for a unit of quality-adjusted consumption is P_{ft}/φ_{ft}. However, for multiproduct firms, the formula for P_{ft} depends on the demand system, and therefore the formula for a firm's real output also depends on the demand system. *In other words, for multiproduct firms, the concept of real output is not independent of the demand system, so all attempts to measure productivity based on a real output concept contain an implicit assumption about the structure of the*

[19] The UPC own price elasticity derived in equation (39) is given by

$$\frac{\partial C_{ut}}{P_{ut}}\frac{P_{ut}}{C_{ut}} = (\sigma_F - 1)\, S_{ft}S_{ut} + (\sigma_U - \sigma_F)\, S_{ut} - \sigma_U.$$

demand system.[20]

This problem is more pernicious than simply saying that firm prices are measured with error; the errors are likely to be systematic. Equation (33) indicates that, in the CES system, the use of average goods prices to measure firm-level prices will overstate the price level (and understate real output) more for large multiproduct firms than for small single-product ones. Therefore, if the true model is CES but a researcher uses a quantity-weighted average of the prices of the goods supplied by the firm to measure firm prices—what we term a "conventional" price index or P_{ft}^{Conv}—the results are likely to underestimate productivity differences.

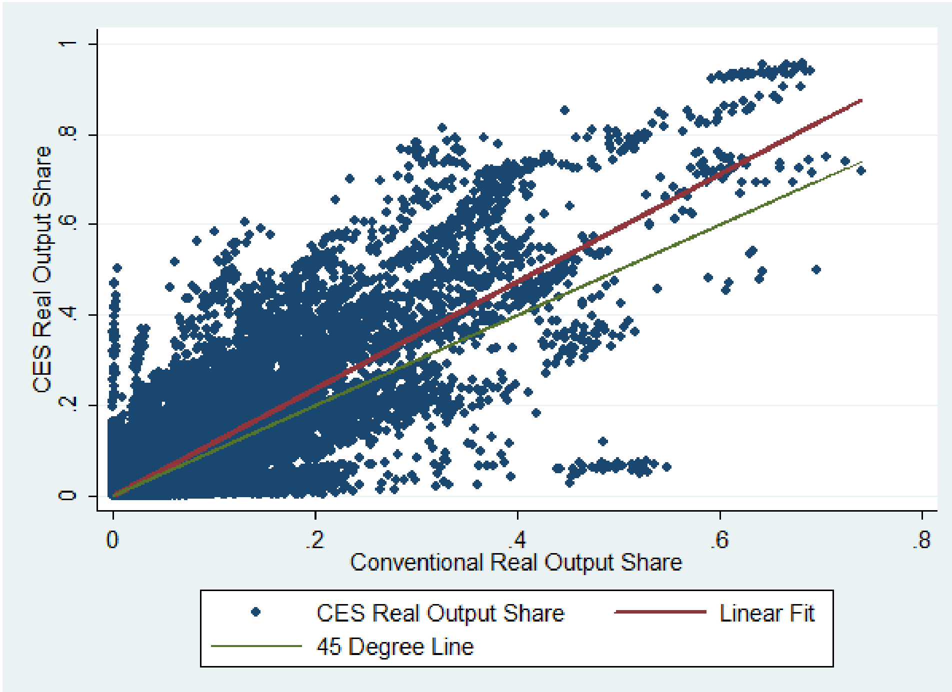
But how big is this problem in practice? If we denote the consumer's expenditure on a firm's output by E_{ft}, we can write the conventional measure of real output, Q_{ft}^{Conv}, as E_{ft}/P_{ft}^{Conv}, and the CES measure of real output, Q_{ft}^{CES}, as $E_{ft}/\left(P_{ft}/\varphi_{ft}\right)$. These measures will vary across sectors due to the units, so in order to compare size variation across sectors we will work with unitless shares of each firm in total output (*i.e.,* $Q_{ft}^{CES}/\sum_{f\in g} Q_{ft}^{CES}$ and $Q_{ft}^{Conv}/\sum_{f\in g} Q_{ft}^{Conv}$). Figure 1 plots these two measures of real output. As one can see in Figure 1, the choice of price index matters enormously for the computation of real output. Moreover, our estimate of σ_U suggests that there are large downward biases in the measurement of real output of large firms. While, on average, the conventional measure of real output is quite close to the CES measure for small firms it systematically understates the value of real output for firms with large market shares. We can econometrically measure the magnitude of this bias by regressing $\ln\left(P_{ft}^{Conv}/\left(P_{ft}/\varphi_{ft}\right)\right)$ on $\ln E_{ft}$ (with product group fixed effects) and examining the coefficient on log sales. If there were no bias, we would expect a coefficient of zero, but we actually obtain a coefficient of 0.3447 (s.e. 0.0004, $R^2 = 0.3$) indicating that failing to take into account the multiproduct nature of firms causes us to significantly underestimate the size of large firms. This estimate is economically significant as well. It implies that holding fixed a firm's conventional price index, every one percent increase in firm sales is associated with a 1.3 percent increase in its real output in a CES model. The difference between the two estimates reflects the importance of assumptions about demand when constructing a firm price index.

A second striking feature of our estimated results is that the elasticity of substitution among varieties supplied by a firm is always larger than that that between firms (*i.e.,* $\sigma_U > \sigma_F$). It is important to remember that this is not a result that we imposed on the data. Moreover, most of the elasticities are precisely estimated—in 82 percent of the cases, we can statistically reject the hypothesis that $\sigma_U = \sigma_F$ at the 5% level. The higher elasticities of substitution across UPCs than across firms imply that varieties are more substitutable within firms than across firms, which implies cannibalization effects from the introduction of new varieties by firms. Moreover, the fact that the estimated elasticities of substitution are always greater than one implies that firms' varieties are substitutes, which is required for positive markups of price over marginal costs (*c.f.,* equation (9)).

In order to assess whether our elasticities are plausible, it is useful to compare our estimates with

[20]Since firms supply differentiated products, the relevant measure of productivity is TFPR (Foster, Haltiwanger and Syverson 2008), which necessarily depends on prices and hence the specification of demand.

Figure 1: Bias in Conventional Real Output Share



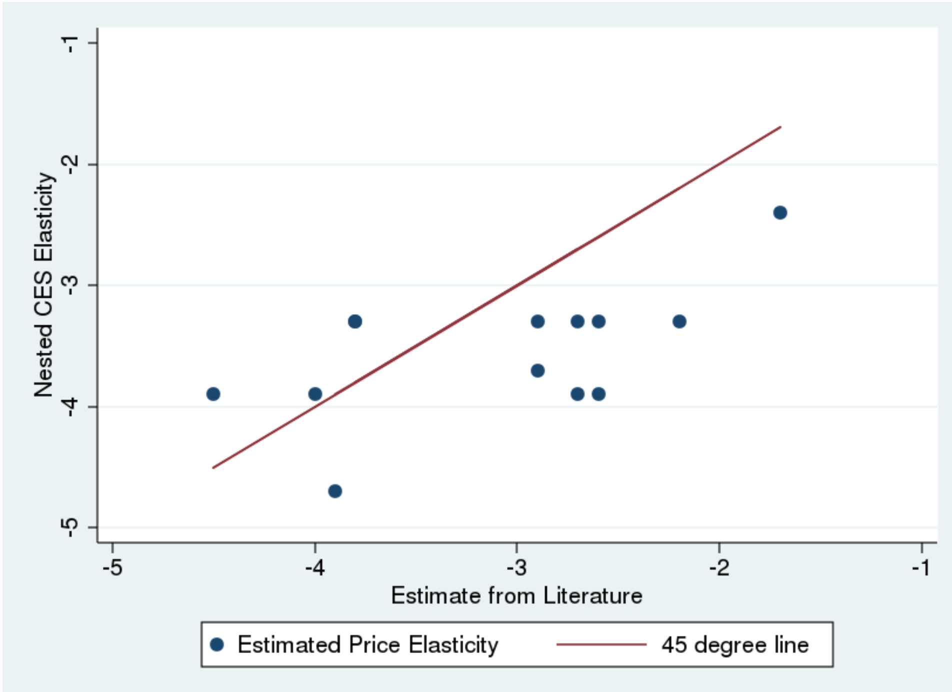
Note: Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

those of other papers. In order to do this, we restricted ourselves to comparing our results with studies that used US scanner data and estimated elasticities for the same product groups as ours. We do this because, as Broda and Weinstein (2006) show, elasticity estimates for aggregate data can look quite different than those for disaggregate data. Unfortunately, we did not find studies estimating the elasticity of substitution within firms, but we did find studies that examined elasticities akin to our cross-firm elasticity, σ_F. Figure 2 plots our estimates of σ_F against other estimates from Gordon et al. (2013) and the literature review contained in the web appendix to their paper. Although these other studies use different methods, we find that our estimates are very similar with a statistically significant correlation of 0.59 and the points arrayed fairly close to the 45-degree line. Therefore, while our empirical approach has a number of novel features in modeling multiproduct firms that are of positive measure relative to the markets in which they operate, the empirical estimates generated by our procedure are reasonable compared to the benchmark of findings from other empirical studies.

6.2 Quality

The quality parameters are a second key component of our decomposition. Our estimation procedure allows us to estimate a different φ_{ut} for every quarter in our dataset. We have strong priors that the quality of each UPC should be fairly stable across time. One way to gauge this stability is to regress these quality parameters on UPC fixed effects in order to determine how much of a UPC's quality is

Figure 2: Price Elasticities Compared to Literature



Note: Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

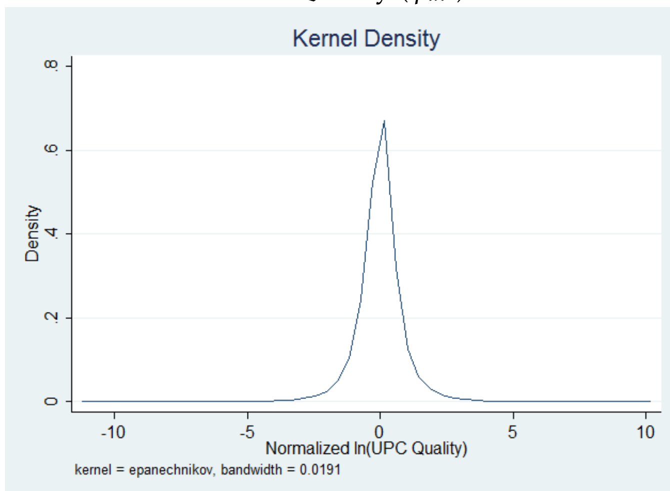
common across time periods. When we do this and include time fixed effects to control for inflation and other common demand shocks, we find that the R^2 is 0.84, which implies that very little of the variation in our quality measures arise from changes in UPC quality. We should also expect firm average quality, φ_{ft}, to exhibit a strong firm component but to be less stable over time. This is exactly what the data reveals. Running the same regression for firm average quality, we find that 83 percent of the variance can be explained by firm fixed effects. Consistent with these results, we also estimate that firm average quality exhibits more variation overall (not just across time) than UPC quality or marginal cost. This can be seen in Figure 3, which plots kernel density estimates of UPC quality, firm average quality, and marginal cost.

6.3 Markups

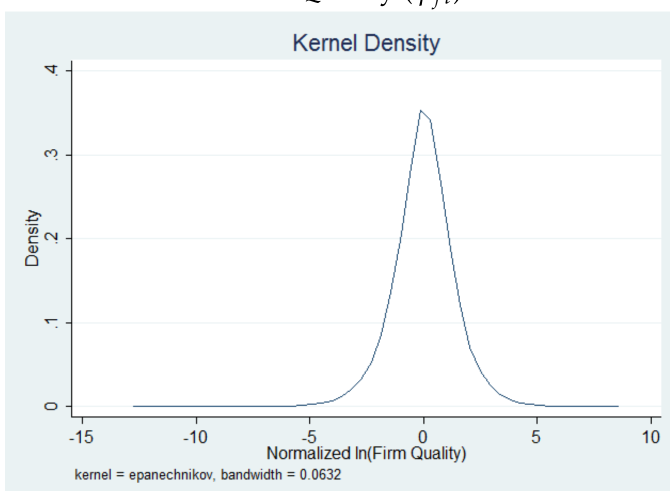
Using our estimated elasticities of substitution, we can compute implied firm markups. As discussed above, multiproduct firms internalize the fact that they supply a firm consumption index and hence choose a price for each UPC that depends on the perceived elasticity of demand for the firm consumption index as a whole (ε_{ft}). Although demand exhibits a constant elasticity of substitution, this perceived elasticity of demand differs across firms because they internalize the effect of their pricing decisions on market price indices. For any firm with a strictly positive expenditure share, the perceived elasticity of demand (ε_{ft}) is strictly less than consumers' elasticity of substitution across varieties ($\varepsilon_{ft} < \sigma_F$). Larger firms that account for greater shares of expenditure have lower per-

Figure 3: Kernel Densities of Quality and Cost Estimates

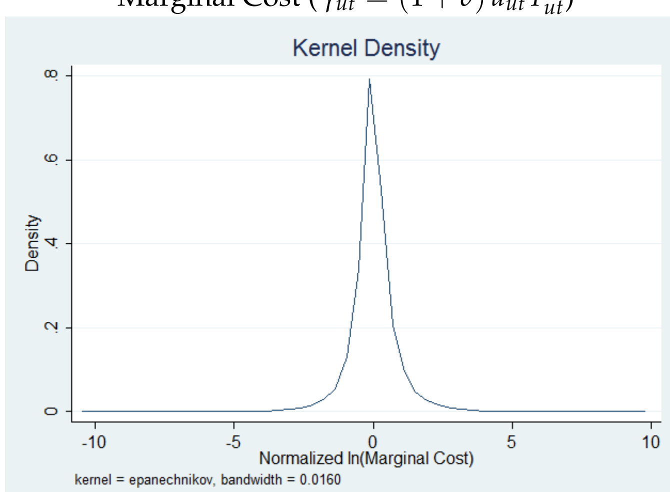
UPC Quality (φ_{ut})



Firm Quality (φ_{ft})



Marginal Cost ($\gamma_{ut} = (1+\delta)\, a_{ut} Y_{ut}^{\delta}$)



Note: The UPC quality is measured by $\ln \varphi_{ut}$. Firm quality is measured by$\ln \varphi_{ft}$, and marginal cost is computed as $\ln\left(\gamma_{ut}/\tilde{\gamma}_{ft}\right)$. Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

ceived elasticities of demand and hence charge higher markups of price over marginal costs (μ_{ft}), as summarized in equations (7) and (8) that are reproduced below:

$$\mu_{ft} = \frac{\varepsilon_{ft}}{\varepsilon_{ft} - 1},$$
$$\varepsilon_{ft} = \sigma_F - (\sigma_F - 1)\, S_{ft}.$$

There are two important dimensions across which markups vary. The first is cross-sector variation with captures the fact that σ_F varies systematically across product groups, and the second is within-sector variation which captures the fact that larger firms have higher markups within a sector than smaller firms. As discussed above, our procedure for estimating the elasticities $\{\sigma_U, \sigma_F\}$ is robust to the assumption of either price or quantity competition, because the firm markup differences out in (27) and we observe prices. However, the markup formula, and hence the decomposition of prices into markups and marginal cost, depends in an important way on the form of competition. In our baseline specification, we assume Bertrand competition in prices, but we report the derivation of the markup under Cournot competition in quantities in the appendix. A key takeaway from this comparison is that Cournot competition generates substantially higher markups for larger firms than Bertrand competition if goods are substitutes. Additionally, as shown in the appendix, under Bertrand competition, the relative markups of firms depend solely on the distribution of market shares (and hence are robust to different estimates of the elasticity of substitution between firms). In contrast, under Cournot competition, the relative markups of firms depend on both the distribution of market shares and the elasticity of substitution between firms.

Table 6 presents data on the distribution of markups across all firms in our 100 product groups. The median markup is 33 percent, which is slightly lower than Domowitz et al. (1988) estimate of 36 percent for U.S. consumer goods, and slightly above the median markups estimated by De Loecker and Warzynski (2012) for Slovenian data which range from 17 to 28 percent. This suggests that our markup estimates are reasonable in the sense that they do not differ greatly from those found in prior work.

Table 6: Distribution of Markups

Ranked Percentile	Cournot Using ε_F	Bertrand Using ε_F	Using σ_F	Using σ_U
10	0.57	0.57	0.57	0.24
25	0.45	0.45	0.45	0.23
50	0.33	0.33	0.33	0.17
75	0.24	0.24	0.24	0.14
90	0.18	0.17	0.16	0.09

Note: Markup=(Price-Marginal Cost)/Marginal Cost. Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

While the markups of most firms are essentially the same regardless of whether we assume firms compete through prices or quantities, the markups of the largest firms are substantially different. Table 7 shows the distribution of the markups for the largest and second largest firms by product group. In each case, we report the markup relative to the average markup in the product group, so that one can see how different the markup of the largest firm is. The median largest firm within product groups has a markup that is 24 percent larger than average if we assume competition is Bertrand and almost double the average if we assume competition is Cournot. However, as one can also see from this table, these markups drop off quite rapidly for the second largest firm. Therefore, both modes of competition suggest that very few firms can exploit their market power.

Table 7: Distribution of Markups Relative to Product Group Average

	Cournot		Bertrand	
Ranked Percentile	Largest Firm	Second Largest	Largest Firm	Second Largest
10	4.68	2.59	1.73	1.28
25	2.98	1.72	1.41	1.18
50	1.94	1.49	1.24	1.13
75	1.54	1.32	1.16	1.09
90	1.37	1.23	1.10	1.06

Note: Percentiles are decreasing: largest is ranked first. Markup=(Price-Marginal Cost)/Marginal Cost. Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

6.4 Cannibalization Rate

Our estimated elasticities of substitution also determine the magnitude of cannibalization effects in the model. As discussed above, when a multiproduct firm chooses whether to introduce a new variety, it takes into account the impact of this product introduction on the sales of its existing varieties. The partial elasticity of the revenue of existing UPCs with respect to the introduction of a UPC is given by (13), which is reproduced below:

$$\text{Cannibalization rate} \equiv \left[\left(\frac{\sigma_U - \sigma_F}{\sigma_U - 1} \right) + \left(\frac{\sigma_F - 1}{\sigma_U - 1} \right) S_{ft} \right] s_{N_{ft}} N_{ft}.$$

Since this cannibalization rate is defined as the partial elasticity of the sales of existing products with respect to the number of products, and can be expressed in terms of observed expenditure shares, it takes the same value under both price and quantity competition, as shown in the appendix.

As we mentioned earlier, a cannibalization level of 0 corresponds to equal differentiation of a firm's products from its rivals and a level of 1 corresponds to a world in which the products of firms are perfect substitutes. In Table 8, we present the average estimated markups and cannibalization levels for firms by decile of size. As one can see from the table, these are remarkably stable, which reflects the fact that even in the upper decile of firms most firms have trivial market shares ($S_f \approx 0$). Interestingly, we see that cannibalization levels lie almost exactly between the benchmarks of perfect substitutes and equal differentiation. This cannibalization level implies that about half of the sales of a new product introduced by a firm comes from the sales of existing products and half from the sales of other firms. The fact that these cannibalization rates are much less than one, however, underscores the fact that it is not appropriate to treat the goods supplied by a multiproduct firm as perfect substitutes nor is it appropriate to treat multiproduct firms as if the introduction of a new good had no impact on the sales of existing goods.

Table 8: Markups and Cannibalization by Decile

Ranked Decile	Average Cournot Markup	Average Bertrand Markup	Average Cannibalization Rate
1	0.40	0.38	0.51
2	0.38	0.37	0.50
3	0.37	0.37	0.50
4	0.37	0.37	0.50
5	0.37	0.37	0.50
6	0.37	0.37	0.50
7	0.37	0.37	0.50
8	0.37	0.37	0.50
9	0.37	0.37	0.50
10	0.37	0.37	0.50

Note: Largest decile is ranked first. Weighted average across product groups. Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

Moreover, there is a slight tendency for the markups and cannibalization levels to rise with firm size reflecting the fact that large firms have higher markups and take into account the fact that more of their sales of new products comes at the expense of their existing products. We can see this more clearly in Table 9, where we focus on the ten largest firms in the sector. We estimate that the markup of the largest firm is on average 32 and 120 percent larger than the markups of firms in the first nine deciles under price and quantity competition respectively (comparing Tables 9 and 8), but the markup of the fifth-largest firm is not very different. Thus, the ability to exploit market power is a feature of very few firms on average. Similarly, cannibalization levels tend to rise for the the very largest firms. We estimate that when the largest firm in a sector, which has an average market share of 23 percent, introduces a new product, 62 percent of the sales of that product comes from from the sales of its existing products. Since we saw from Table 3 that the largest firms typically have the most products per firm, this implies that cannibalization is likely to be a first-order issue for them.

Table 9: Markups and Cannibalization by Firm Rank

Firm Rank	Cournot Markup	Bertrand Markup	Cannibalization Rate
1	0.826	0.496	0.617
2	0.576	0.428	0.565
3	0.498	0.407	0.544
4	0.455	0.397	0.531
5	0.437	0.392	0.525
6	0.424	0.388	0.521
7	0.415	0.385	0.518
8	0.413	0.384	0.517
9	0.410	0.383	0.516
10	0.403	0.382	0.514

Note: Largest firm is ranked first. Weighted averages across product groups. Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

6.5 Decomposing Firm Sales

The first panel of Table 10 presents the decompositions described in equations (19) to (23) for the full sample of firms and the second presents the results for the largest firms (those with a market share in excess of 0.5 percent). We run the results over both sets of firms to see if there are any differences between them. Overall, the results suggest a powerful role for quality in accounting for the size of firms. The variance decomposition indicates that 76 percent of the overall size distribution can be attributed to average quality and 21 percent is attributed to scope. Thus, total firm quality accounts for 97 percent of variation in firm sales. If we restrict ourselves to looking at the data for the fifty largest firms we still find that firm average quality explains on average 54 percent of the variation with scope accounting for 26 percent. In other words, even if we restrict our sample to the largest firms, quality variation accounts for 80 percent of the variation across firms.

What is perhaps most surprising is the small role played by cost in the determination of firm size. There are two important features of this decomposition to bear in mind. First, a positive coefficient means that larger firms have *lower* costs: *i.e.*, cost differentials help explain size differentials. Second, our decomposition allows us to split firm-cost differentials into two components: average marginal cost differentials and the dispersion of costs across products ("cost dispersion"). In all cross-sectional decompositions, the cost dispersion term is positively correlated with firm size. The intuition is simple. Larger firms tend to sell more products and therefore are better positioned to benefit from cost differentials in supplying their consumption index. The results for average marginal costs are more mixed. In the full sample, we find almost no association between firm size and average costs indicating that larger firms have essentially the same average marginal costs as smaller firms. Their total costs are lower because they can take advantage of the cost dispersion in their output. Interestingly, we find a much stronger association with average costs when we restrict our sample to only the largest fifty firms. Here, we find that 18 percent of the firm size distribution can be attributed to lower average costs of large firms and a quarter of of the overall distribution due to these cost differences. Moreover, the results lend support to the managerial competence approach to thinking about cost differences—large firms tend to have higher quality *and* lower costs. Finally markup variation, which depends on both quality and cost through expenditure shares, plays a very small role in understanding the firm size distribution.[21]

Figure 4 presents the results of the cross-sectional decomposition for the largest 50 firms in each product group. Each point represents the average contribution of a particular factor towards understanding the relative level of sales in a sector. For example, the column of points at the right of the figure indicate that the largest firm in a product group is on average almost nine log units larger than the average firm. Of this nine-log-unit difference the diamond-shaped point indicates that on average about 4.9 log units (54 percent) can be attributed to average quality differences. The circular point reveals that about 2.3 log units (25 percent) is on average attributable to scope differences,

[21]Consistent with these findings, Roberts et al. (2011) find substantial heterogeneity in demand across Chinese footwear producers, while Foster, Haltiwanger and Syverson (2013) emphasize the role of demand in the growth of entering firms.

Table 10: Variance Decompositions

Decomposition		Quality	Scope	Average MC	Cost Dispersion	Markup	Upgrading
PANEL A: ALL FIRMS							
Cross-Sectional	Coefficient	0.758	0.2090	−0.036	0.0707	−0.00195	
	Std.Error	0.002	0.0002	0.002	0.0001	0.00001	
Firm Growth	Coefficient	0.905	0.1571	−0.131	0.0687	−0.00059	0.0003
	Std.Error	0.009	0.0012	0.008	0.0010	0.00004	0.0014
PANEL B: TOP 50 FIRMS							
Cross-Sectional	Coefficient	0.544	0.2637	0.175	0.0613	−0.04436	
	Std.Error	0.013	0.0016	0.013	0.0008	0.00016	
Firm Growth	Coefficient	0.902	0.2597	−0.163	0.0278	−0.02619	0.0233
	Std.Error	0.053	0.0079	0.051	0.0082	0.00131	0.0138

Note: MC is marginal cost Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

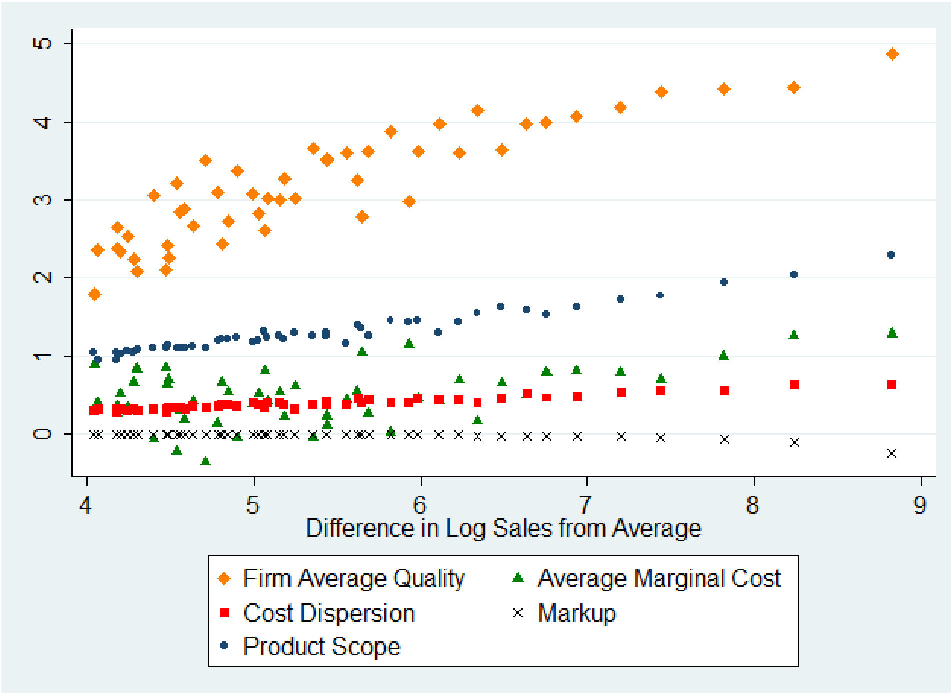
while the triangular and square points indicate that lower average costs and greater cost dispersion account for about 1.3 (14 percent) and 0.6 log units (7 percent) of the sales. These percentages are extremely similar to the percentages for larger firms that we saw in Table 10. However, the fact that each set of points is relatively tightly arrayed along a line indicates that the results presented in Table 10 are not the result of any outlier but are a robust feature of the data.

There are two other features of the data that this plot also makes clear. First, we can also see that while there is a tendency for average marginal costs to fall with size for the four largest firms, which accounts for the upward trend in the sequence of triangular points, this upward trend is not a robust feature of the data. This pattern helps explain why we found in Table 10 that average marginal costs did not fall with firm size when we looked at the full sample of firms, but did fall when we restricted our attention to the very largest firms. Moreover, we can see that the gains from increasing cost dispersion are a very robust feature of the data.

A second feature of the data that suggests welfare implications of the ability of the largest firms to exploit market power is also clearly seen in Figure 4. We saw in Table 10 that while markup variation played almost no role in understanding the firm-size distribution when we focused on the full sample of firms, it was negatively associated with firm sales when we examined the largest firms. Figure 4 shows that this negative association comes from the fact that if it were not for the fact that the very largest two or three firms had higher markups (a result we also saw in Table 9), their output levels would be higher. The implied inefficiently low levels of output of the largest firms suggests that the monopolistic competition model breaks down for these firms and motivates our attempt to quantify the welfare impact of this pattern in the data in Section 7.

As we showed in Section 4.6, we can also examine the time-series determinants of firm growth by differencing with respect to time (*i.e.*, using $\Delta^{g,t}$ instead of Δ^{g} in the decompositions). As we can see in Table 10, the importance of quality rises when understanding firm growth. Virtually all of a

Figure 4: Sales Decomposition by Firm Rank

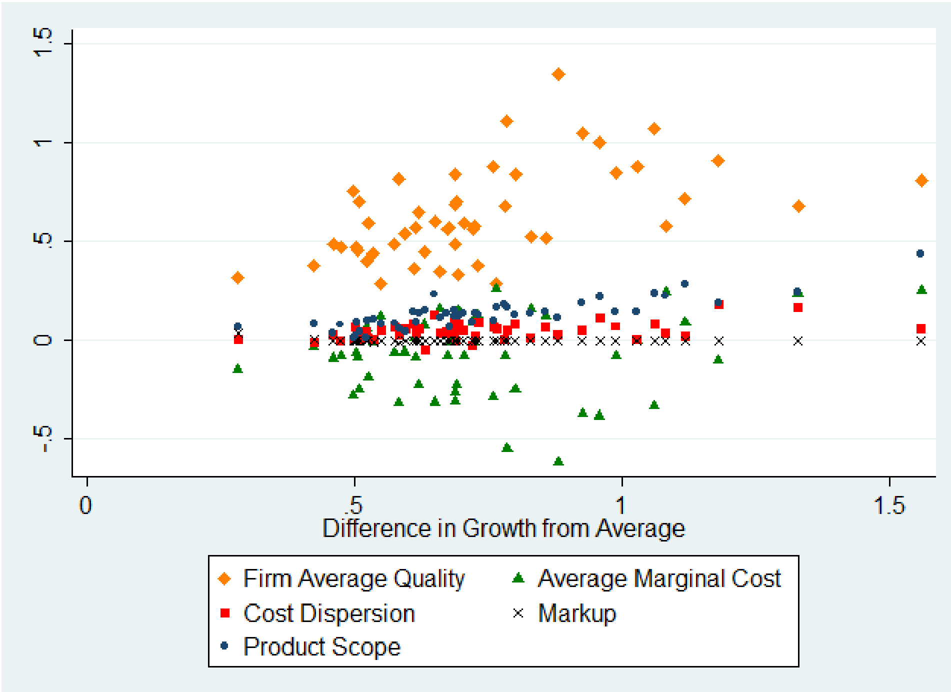


Note: The graph contains the weighted average across product groups of the decomposition of firm sales into firm average quality, marginal cost, markup and product scope for the fifty firms with the largest average market share in 2004 and 2011. The vertical axis measures the contribution of each factor given by the left-hand sides of equations (19) to (23). The horizontal axis is the log difference between the firm's sales and that of the average firm defined by the right-hand side of these equations. Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

firm's growth can be understood in terms of increases in quality and scope with cost changes playing almost no role. Interestingly, quality upgrading—*i.e.*, the addition of new higher quality products—plays almost no role in the typical firm's expansion. Most of the expansion comes from shifts in demand towards the firm's existing products, which may reflect marketing and other taste shifts.

Figure 5 presents the results from this decomposition for the fifty firms whose average market shares in 2004 and 2011 were the highest. Thus, the column of points on the right of the figure indicates that on average the fastest growing firm in a product group on average had sales growth that was 1.6 log units faster than the average firm (or about 26 percent per year). Of this faster growth, the diamond-shaped point indicates that close to one log unit (about 50 percent) was due to quality improvements, with most of the remainder due to scope changes. The points in these scatters are a bit less tightly distributed, indicating that there is more volatility in the estimated determinants of firm growth than in firm shares, but the results consistently show that the two main determinants of firm growth are firm average quality and product scope. These results are also quite similar to to those we saw in Table 10, with now the new piece of information being that we can see that the earlier results are not being driven by any particular firm or set of firms.

Figure 5: Growth Decomposition the Fifty Largest Firms



Note: The graph contains the weighted average across product groups of the growth decomposition of the growth rate in firm sales into estimated firm average quality, marginal cost, and product scope for the top fifty firms by average market share in 2004 and 2011. The vertical axis measures the contribution of each factor given by the left-hand sides of equations (19) to (23). The horizontal axis is the log difference between the firm's sales and that of the average firm growth rate defined by the right-hand side of these equations. Calculated based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

Taken together, the decomposition results are quite striking in their consistency about the sources of firm heterogeneity. Regardless of whether we examine firms in the cross-section or in the time-series, improvements in firm average quality followed by firm scope are the most important drivers of firm sales for both small and large firms alike. The role played by average cost differences tends to be small and depend on the sample and decomposition method used. But we consistently find a positive contribution towards differences in firm size from the cost dispersion term. Finally, we see that although markup differences are unimportant determinants of firm sales for most firms, they do appear to be important for the very largest firms something that suggests aggregate welfare implications from the exploitation of market power by these firms.

6.6 Robustness

One limitation of our approach is that to structurally estimate the model we need to assume a particular nesting structure for demand consisting of firms and products within firms. We choose this nesting structure to connect with the existing literature on measuring firm productivity (which implicitly or explicitly treats firms as a nest) and as a natural approach to modeling multi-product firms. But it is reasonable to wonder whether our results are robust to other nesting structures. In this section, we undertake a robustness test, in which we demonstrate that our main finding of the importance of quality in explaining sales variation is not sensitive to our assumed nesting structure. To do so, we use the CES functional form to derive a reduced-form equation that can be estimated for alternative nesting structures. Although this reduced-form equation cannot be used to recover structural parameters, it can be used to provide a lower bound on the importance of quality variation for alternative nesting structures.

To provide evidence on the role of quality variation under alternative nesting structures, we combine the following two results for CES demand. First, from (31), double-differenced expenditure shares depend on double-differenced price indices for any tier of utility j. Second, from (33), the price index for one tier of utility j can be expressed in terms of expenditure shares and the geometric mean of prices for the lower tier of utility k. Combining these results, and reintroducing the product group subscript, we obtain the following reduced-form equation that can be estimated using OLS:

$$\triangle^{j,t}_{-} \ln S_{jgt} = \alpha_{jg} + \beta_{1jg} \triangle^{j,t}_{-} \ln \tilde{P}_{jt} + \beta_{2jg} \triangle^{j,t}_{-} \ln \left[\sum_{u \in b} \frac{S_{kgt}}{\tilde{S}_{jgt}} \right] + \epsilon_{jgt}, \tag{35}$$

where $\triangle^{j,t}_{-}$ is again the double-difference operator; $\widetilde{P}_{jt} = \exp \left\{ \frac{1}{N_{jt}} \sum_{k=1}^{N_{jt}} \ln P_{kt} \right\}$ and the error term ϵ_{jgt} captures double-differenced quality.

In our structural model, the coefficient β_{2jg} can be related to structural parameters of the model ($\beta_{2jg} = 1/\left(\sigma_{kg} - 1\right)$). In the reduced-form regression (35), this coefficient does not have a structural interpretation, because quality for tier j (ϵ_{jgt}) can be correlated with the geometric mean of prices in the lower tier k ($\widetilde{P}_{jgt}$). Therefore, if the reduced-form regression (35) is estimated using OLS, the coefficient β_{2jg} is subject to omitted variable bias and is not consistently estimated. For this reason, we use our structural approach developed above to estimate the elasticities of substitution and elasticity of marginal cost $\{\sigma_{Ug}, \sigma_{Fg}, \delta_g\}$ for our assumed nesting structure.

Nonetheless, the reduced-form specification (35) can be used to provide a lower bound on the importance of quality for alternative nesting structures. The reason that this specification provides a lower bound is that quality is captured in the estimated residual ($\widehat{\epsilon}_{jgt}$). Therefore, the component of quality that is correlated with the explanatory variables is attributed to the coefficients $\{\beta_{1jg}, \beta_{2jg}\}$ on these explanatory variables, leaving only the orthogonal component of quality in the estimated residual ($\widehat{\epsilon}_{jgt}$).

In Table 11, we report the average R^2 across all product groups (g) from estimating this reduced-form regression for alternative nesting structures, where one minus this R^2 corresponds to the variance in double-differenced expenditure shares that is unexplained by the explanatory variables (and is instead explained by quality). We estimate the regression separately for each product group and year, and summarize the mean and standard deviation of the R^2, weighting each product group and year by their sales.

In the first row of the table, we report results for our assumed nesting structure. In the remaining rows of the table, we report results for a wide range of alternative nesting structures. Some of these alternative nesting structures disaggregate firms relative to our baseline specification (e.g. Specification II uses Product Group and Brand-Module instead of Product Group and Firm, where Brand Module distinguishes between different brands supplied by a firm, such as Coke and Sprite soda for the Coca-Cola corporation). Other alternative nesting structures omit firms as a nest altogether (e.g. Specification IV uses Module and Brand Module, where Module corresponds to different types of products within a product group, such as diet and regular soda). Across this wide range of specifications, we find that quality accounts for at least 43 percent of the variation in double-differenced sales. Thus, regardless of one's choice of nesting structure, it is hard to escape the conclusion that around half of sales variation is attributable to quality variation. Since this lower bound also leaves out the important role played by scope on demand, this further reinforces the importance of overall quality (including scope) in explaining sales variation.

Table 11: Within Nest R^2

Specification	Nest 1	Nest 2	Nest 3	Nest 4	Mean	Std. Dev.
I	Product Group	Firm	UPC	N/A	0.53	0.05
II	Product Group	Brand-Module	UPC	N/A	0.46	0.07
III	Product Group	Firm-Module	UPC	N/A	0.47	0.07
V	Module	Firm-Module	UPC	N/A	0.48	0.06
V	Module	Firm-Module	UPC	N/A	0.49	0.05
VI	Module	Firm-Module	Brand	UPC	0.57	0.07

Note: Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

7 Counterfactuals

We found in Table 9 that the monopolistic competition benchmark breaks down substantially for the largest firms. Moreover, we learned from Figure 4 that the estimated exploitation of market power by the largest firms implies reductions in these firms' sales. We now use the structure of the model to undertake counterfactuals to assess the quantitative relevance of our estimated departures from monopolistic competition for aggregate welfare. We consider two types of counterfactuals. First, we consider the implications for consumers from a regulator preventing large firms from exercising their market power and thereby requiring each firm to compete under monopolistic competition. Second, we consider the implications of allowing firms to supply a number of differentiated products. Thus, the second type of counterfactuals examines the gains to consumers from the different varieties supplied by each firm.

In each of these counterfactual exercises, we examine the impact on consumers under both Bertrand and Cournot competition. In all cases, we hold constant aggregate expenditure (E_t), the sets of active UPCs and firms $\{U_{ft}, F_{gt}\}$, and the sets of firm qualities, UPC qualities and cost shifters $\{\varphi_{ft}, \varphi_{ut}, a_{ut}\}$. Therefore we abstract from general equilibrium effects through aggregate expenditure, factor prices, and entry and exit. Our goal is not to capture the full general equilibrium effects of each counterfactual, but rather to show that our firm-level estimates are of quantitative relevance for aggregate outcomes such as welfare.

In the first set of counterfactuals (Counterfactuals I-B and I-C), we compare the actual equilibrium in which firms exploit their market power with a counterfactual equilibrium in which a price regulator requires all firms to charge the same constant markup over marginal cost equal to the monopolistically competitive firm markup:

$$\mu_{ft} = \frac{\sigma_F}{\sigma_F - 1}. \tag{36}$$

Moving from the actual to the counterfactual markup reduces the relative prices of successful firms with high market shares and hence redistributes market shares towards these firms. Therefore, we expect these counterfactuals to increase welfare and the dispersion of firm sales, but the magnitude of these effects depends on the estimated parameters $\{\sigma_U,\ \sigma_F,\ \delta\}$ and the sets of firm qualities, product qualities and cost shocks $\{\varphi_{ft},\ \varphi_{ut},\ a_{ut}\}$.

Given the assumed constant markup in equation (36), we solve for the counterfactual equilibrium using an iterative procedure to solve for a fixed point in a system of five equations for expenditure shares and price indices $\{S_{ut},\ P_{ut},\ S_{ft},\ P_{ft},\ P_{gt}\}$. These five equations are the UPC expenditure share (given by equation (4) for $k = u$), the UPC pricing rule (equation (9)), the firm expenditure share (given by equation (4) for $k = f$), the firm price index (given by (3) for $k = f$) and the product group price index (given by equation (3) for $k = g$). We use the resulting solutions for product group price indices together with aggregate expenditure to compute counterfactual aggregate welfare:

$$W_t = \frac{E_t}{\prod_{g=1}^{G} P_{gt}^{\varphi_{gt}}}. \tag{37}$$

In Counterfactuals II-B and II-C, we examine the quantitative relevance of multiproduct firms. We compare the actual equilibrium in which firms supply multiple products to a counterfactual equilibrium in which firms are restricted to supply a single UPC (their largest). Counterfactual II-B undertakes this comparison for the actual equilibrium under Bertrand competition, while Counterfactual II-C undertakes the same comparison for the actual equilibrium under Cournot competition. We again solve for the counterfactual equilibrium by solving for a fixed point in the system of five equations for expenditure shares and price indices discussed above.

Table 12 reports the results from both sets of counterfactuals. Each row of the table corresponds to a different counterfactual. The second column reports the coefficient of variation of firm sales in the counterfactual relative to that in the actual data. The third column reports aggregate welfare in the counterfactual relative to that in the actual data.

In Counterfactual I-B, we find that moving from Bertrand to monopolistic competition increases aggregate welfare by 4 percent, which is comparable to standard estimates of the welfare gains from trade for an economy such as the United States. In Counterfactual I-C, we find much larger welfare effects of moving from Cournot to monopolistic competition, which increases aggregate welfare by

Table 12: Counterfactual Exercises (Relative to Observed)

	Coef. Variation (Firm Sales)	Aggregate Welfare
I-B: Bertrand to Monopolistic Comp.	1.10	1.04
I-C: Cournot to Monopolistic Comp.	1.24	1.15
II-B: Multiproduct to Single-Product Bertrand	0.82	0.70
II-C: Multiproduct to Single-Product Cournot	0.86	0.70

Note: Each statistic is expressed as the value in the counterfactual relative to the value in the observed data. Therefore relative coefficient variation sales is the coefficient of variation of firm sales in the counterfactual divided by the coefficient of variation in the observed data. Calculations on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

around 15 percent. These much larger welfare effects reflect the fact that large firms charge much higher markups under Cournot than under Bertrand as we saw in Table 7. This result implies larger price reductions from moving to monopolistically competitive markups, and hence larger increases in welfare. In both Counterfactuals I-B and I-C, the relative prices of the largest firms fall and demand is elastic. Therefore, the coefficient of variation of firms sales increases, by around 10 percent for Bertrand competition and around 24 percent for Cournot competition. Again the larger effects for Cournot reflect the greater variation in markups under this mode of competition.

In both Counterfactuals II-B and II-C, we find quantitatively large effects of multiproduct firms. Restricting firms to supply a single product under Bertrand competition reduces aggregate welfare by around 30 percent and reduces the coefficient of variation of firm sales by around 18 percent. The corresponding reductions in aggregate welfare and the coefficient of variation of firm sales under Cournot competition are similar: 30 percent and 14 percent respectively. These results suggest that consumers benefit enormously from the range of products supplied within a single firm.

All of these counterfactuals are subject to a number of caveats. In particular, they do not allow for firm entry or exit, because they hold the set of firms constant, and they abstract from general equilibrium effects by holding aggregate expenditure and factor prices constant. Nonetheless, these counterfactuals establish the quantitative relevance of our departures from the case of atomistic monopolistic competition with single product firms.

8 Conclusions

We develop and structurally estimate a model of heterogeneous multiproduct firms that can be used to decompose the firm-size distribution into the contributions of costs, quality, scope and markups. Our framework requires only price and expenditure data and hence is widely applicable. We use these price and expenditure data to estimate the key parameters of the model, namely the elasticities of substitution between and within firms, and the elasticities of marginal cost with respect to output. We use the resulting parameter estimates and the structure of the model to recover overall firm average quality, the relative quality of firm products, marginal costs, the number of products a firm supplies and the firm's markup.

Our results point to quality differences as being the principal reason why some firms are large and others are not. Depending on the specification considered, we find that 50-75 percent of the variance in firm size can be attributed to quality differences, about 23-30 percent to differences in product scope, and less than 20 percent to average marginal cost differences. If we use a broad measure of total firm quality, which encompasses both average quality and scope, we find that average quality accounts for almost all of firm size differences.

We estimate substantially higher elasticities of substitution between varieties within firms than between firms (median elasticities of 6.9 and 4.3 respectively), implying that a firm's introduction of new product varieties cannibalizes the sales of existing varieties. We estimate that the cannibalization rate for the typical firm is 0.45, roughly half-way in-between the extreme of no cannibalization (equal elasticities of substitution within and between atomistic firms), and the extreme of complete cannibalization (varieties perfectly substitutable within firms).

These findings that firms supply multiple imperfectly substitutable varieties have important implications for the measurement of firm productivity, highlighting the role of assumptions about demand in the measurement of productivity for multiproduct firms. Conventional price indices based on a weighted average of firm prices do not take into account that the true ideal price index for the firm depends on the number of products it supplies. This introduces a systematic bias into the measurement of firm productivity, because larger firms supply more products than smaller firms. We find this bias to be quantitatively large. An increase in firm sales is associated with around a one third larger increase in true real output (using the ideal price index) than in measured real output (using the conventional price index).

We find that most firms charge markups close to the monopolistic competition benchmark of constant markups, because most firms have trivial market shares, and hence are unable to exploit their market power. However, the largest firms that account for up to around 25 percent of sales within sectors charge markups between 32 and 120 percent higher than the median firm. Using the estimated model to undertake counterfactuals, we find that these departures from the monopolistically competitive benchmark have quantitatively relevant effects on aggregate welfare.

References

[1] Agur, Itai (2010) "Trade Liberalization, Firm Selection and Variety Growth," *Review of International Economics*, 18(3), 582-594.

[2] Allanson, Paul and Catia Montagna (2005) "Multiproduct Firms and Market Structure: An Explorative Application to the Product Life Cycle," *International Journal of Industrial Organization*, 23, 587-597.

[3] Amiti, Mary, Oleg Itskhoki and Jozef Konings (2014) "Importers, Exporters and Exchange Rate Disconnect," *American Economic Review*, forthcoming.

[4] Anderson, Simon P., André de Palma and Jacques-François Thisse (1992) *Discrete Choice Theory of Product Differentiation*, Cambridge MA: MIT Press.

[5] Arkolakis, Costas, Arnaud Costinot and Andres Rodriguez-Clare (2012a) "New Trade Models, Same Old Gains?" *American Economic Review*, 102(1), 94-130.

[6] Arkolakis, Costas, Arnaud Costinot and Andres Rodriguez-Clare (2012b) "The Elusive Pro-Competitive Effects of Trade," MIT mimeograph.

[7] Arkolakis, Costas and Marc-Andreas Muendler (2010) "The Extensive Margin of Exporting Products: A Firm-Level Analysis," *NBER Working Paper*, 16641.

[8] Atkeson, Andrew and Ariel Burstein (2008) "Pricing to Market, Trade Costs and International Relative Prices," *American Economic Review*, 98(5), 1998-2031.

[9] Bernard, Andrew B., J. Bradford Jensen and Peter K. Schott (2009) "Importers, Exporters, and Multinationals: A Portrait of Firms in the U.S. that Trade Goods," in *Producer dynamics: New Evidence from Micro Data*, (Eds) T Dunne, JB Jensen and MJ Roberts, Chicago: University of Chicago Press.

[10] Bernard, Andrew B., Stephen J. Redding and Peter K. Schott (2010) "Multiproduct Firms and Product Switching," *American Economic Review*, 100(1), 70-97.

[11] Bernard, Andrew B., Stephen J. Redding and Peter K. Schott (2011) "Multiproduct Firms and Trade Liberalization,"*Quarterly Journal of Economics*, 126(3), 1271-1318.

[12] Berry, Steven T. (1994) "Estimating Discrete-choice Models of Product Differentiation," *RAND Journal of Economics*, 25, 242-262.

[13] Broda, Christian and David E. Weinstein (2006) "Globalization and the Gains from Variety," *Quarterly Journal of Economics*, 121(2), 541-86.

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

[15] Chari, V. V., Patrick J. Kehoe and Ellen R. McGrattan (2007) "Business Cycle Accounting," *Econometrica*, 75(3), 781-836.

[16] Chevalier, Judith A., Anil K. Kashyap, and Peter E. Rossi (2003) "Why Don't Prices Rise during Periods of Peak Demand? Evidence from Scanner Data," *American Economic Review*, 93(1), 15-37.

[17] Das M, Mark J. Roberts, James R. Tybout (2007) "Market Entry Costs, Producer Heterogeneity and Export Dynamics," *Econometrica*, 75(3), 837-73.

[18] De Loecker, Jan and Frederic Warzynski (2012) "Markups and Firm-Level Export Status," *American Economic Review* 2012, 102(6): 2437–2471.

[19] De Loecker, Jan, Penny Goldberg, Amit Khandelwal and Nina Pavcnik (2014) "Prices, Markups and Trade Reform," Princeton University, mimeo.

[20] Dhingra, Swati (2013) "Trading Away Wide Brands for Cheap Brands," *American Economic Review*, 103(6), 2554-84.

[21] Di Giovanni, Julian and Andrei Levchenko (2012) "Country Size, International Trade, and Aggregate Fluctuations in Granular Economies," *Journal of Political Economy*, forthcoming.

[22] Domowitz, Ian, R. Glenn Hubbard, and Bruce C. Petersen (1988), "Business Cycles and the Relationship between Concentration and Price-Cost Margins," *The RAND Journal of Economics*, 17(1), 1-17.

[23] Eckel, Carsten, Leonardo Iacavone, Beata Javorcik and J. Peter Neary (2013) "Multiproduct Firms at Home and Away: Cost Versus Quality-Based Competence," University of Oxford, mimeograph.

[24] Eckel, Carsten and J. Peter Neary (2010) "Multiproduct Firms and Flexible Manufacturing in the Global Economy," *Review of Economic Studies*, 77(1), 188-217.

[25] Edmond, Chris, Virgiliu Midrigan and Daniel Yi Xu (2012) "Competition, Markups, and the Gains from International Trade," *NBER Working Paper*, 18041.

[26] Feenstra, Robert C. (1994) "New Product Varieties and the Measurement of International Prices," *American Economic Review*, 84(1), 157-177.

[27] Feenstra, Robert C. (2014) "Restoring the Product Variety and Pro-competitive Gains from Trade with Heterogeneous Firms and Bounded Productivity," *NBER Working Paper*, 19833.

[28] Feenstra, Rob and Hong Ma (2008) "Optimal Choice of Product Scope for Multiproduct Firms," in (eds) Elhanan Helpman, Dalia Marin and Thierry Verdier, *The Organization of Firms in a Global Economy*, Cambridge MA: Harvard University Press.

[29] Eslava, Marcela, Cecilia Fieler and Daniel Yi Xu (2014) "Trade, Skills and Quality Upgrading: A Theory with Evidence from Colombia," University of Pennsylvania, mimeograph.

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

[31] Foster, Lucia, John Haltiwanger and Chad Syverson (2013) "The Slow Growth of New Plants: Learning about Demand," Chicago Booth School of Business, mimeograph.

[32] Gabaix, Xavier (2011) "The Granular Origins of Aggregate Fluctuations," *Econometrica*, 79, 733-772.

[33] Gordon, Brett R., Avi Goldfarb, and Yang Li (2013), "Does Price Elasticity Vary with Economic Growth? A Cross-Category Analysis," *Journal of Marketing Research*, Vol. L (Feb.), 4-23.

[34] Handbury, Jessie and David E. Weinstein (forthcoming) "Goods Prices and Availability in Cities," *Review of Economic Studies*.

[35] Johnson, Robert C. (2012) "Trade and Prices with Heterogeneous Firms," *Journal of International Economics*, 86 (1), 43-56.

[36] Khandelwal, Amit K. (2010) "The Long and Short (of) Quality Ladders," *Review of Economic Studies*, 77(4), 1450-1476.

[37] Leontief, Wassily (1929) "Ein Versuch Zur Statistichen Analyse von Angebot und Nachfrage," *Weltwirtschaftliches Archiv*, 30(1), 1-53.

[38] Manova, Kalina and Zhiwei Zhang (2012) "Export Prices across Firms and Destinations," *Quarterly Journal of Economics*, 127, 379-436.

[39] Mayer, Thierry, Marc Melitz and Gianmarco I. P. Ottaviano (2014) "Market Size, Competition, and the Product Mix of Exporters," *American Economic Review*, 104(2), 495-536.

[40] Melitz, Marc J. (2003) "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71, 1695-725.

[41] Melitz, Marc J. and Gianmarco I. P. Ottaviano (2008) "Market Size, Trade, and Productivity," *Review of Economic Studies*, 75(1), 295-316.

[42] Melitz, Marc J. and Stephen J. Redding (2013) "Heterogeneous Firms and Trade," *Handbook of International Economics*, Volume 4, forthcoming.

[43] Nocke, Volker and Stephen R. Yeaple (2006) "Globalization and Endogenous Firm Scope", *NBER Working Paper*, 12322.

[44] Roberts, Mark J., Xiaoyan Fan, Daniel Xu and Shengxing Zhang (2011) "A Structural Model of Demand, Cost, and Export Market Selection for Chinese Footwear Producers," *NBER Working Paper*, 17725.

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

[46] Schott, Peter K. (2004) "Across-product Versus Within-product Specialization in International Trade," *Quarterly Journal of Economics*, 119, 647-678.

[47] Stock, James H., Jonathan H. Wright and Motohiro Yogo (2002) "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20(4), 518-529.

[48] Sutton, John (1986) "Vertical Product Differentiation: Some Basic Themes," *The American Economic Review Papers and Proceedings*, 76(2), 393-398.

[49] Vartia, Yrjo (1976) "Ideal Log-Change Index Numbers," *Scandinavian Journal of Statistics*, 3(3), 121-126.

A Appendix A: Derivations

A.1 Derivation of Equations (7)-(9)

The first-order condition with respect to the price of an individual UPC implies:

$$Y_{ut}+\sum_{k=\underline{u}_{ft}}^{\underline{u}_{ft}+N_{ft}}\left(P_{kt}\frac{dY_{kt}}{dP_{ut}}-\frac{dA_k\left(Y_{kt}\right)}{dY_{kt}}\frac{dY_{kt}}{dP_{ut}}\right)=0. \tag{38}$$

Using equation (5) and setting UPC supply equal to demand we have

$$\frac{\partial Y_{kt}}{\partial P_{ut}}=\left(\sigma_f-1\right)\frac{Y_{kt}}{P_{gt}}\frac{\partial P_{gt}}{\partial P_{ut}}+\left(\sigma_u-\sigma_f\right)\frac{Y_{kt}}{P_{ft}}\frac{\partial P_{ft}}{\partial P_{ut}}-\sigma_u\frac{Y_{kt}}{P_{ut}}\frac{\partial P_{kt}}{\partial P_{ut}}.$$

We now can use equation (4) to solve for the elasticities and rewrite $\frac{\partial Y_k}{\partial P_u}$ as

$$\begin{aligned}\frac{\partial Y_{kt}}{\partial P_{ut}} &= \left(\sigma_f-1\right)\left(\frac{\partial P_{gt}}{\partial P_{ft}}\frac{P_{ft}}{P_{gt}}\right)\left(\frac{\partial P_{ft}}{\partial P_{ut}}\frac{P_{ut}}{P_{ft}}\right)\frac{Y_{kt}}{P_{ut}}+\left(\sigma_u-\sigma_f\right)\left(\frac{\partial P_{ft}}{\partial P_{ut}}\frac{P_{ut}}{P_{ft}}\right)\frac{Y_{kt}}{P_{ut}}-\sigma_u\frac{Y_{kt}}{P_{ut}}1_{\{u=k\}}\\ &= \left(\sigma_f-1\right)S_{ft}S_{ut}\frac{Y_{kt}}{P_{ut}}+\left(\sigma_u-\sigma_f\right)S_{ut}\frac{Y_{kt}}{P_{ut}}-\sigma_u\frac{Y_{kt}}{P_{ut}}1_{\{u=k\}}.\end{aligned} \tag{39}$$

If we now substitute equation (39) into equation (38) and divide both sides by Y_u, we get

$$\begin{aligned}&1+\sum_k\left(\sigma_f-1\right)S_{ft}S_{ut}\frac{P_{kt}Y_{kt}}{P_{ut}Y_{ut}}+\sum_k\left(\sigma_u-\sigma_f\right)S_{ut}\frac{P_{kt}Y_{kt}}{P_{ut}Y_{ut}}-\sigma_u\\ &-\sum_k\left(\sigma_f-1\right)S_{ft}S_{ut}\frac{\frac{\partial A_k(Y_{kt})}{\partial Y_{kt}}Y_{kt}}{P_{ut}Y_{ut}}-\sum_k\left(\sigma_u-\sigma_f\right)S_{ut}\frac{\frac{\partial A_k(Y_{kt})}{\partial Y_{kt}}Y_{kt}}{P_{ut}Y_{ut}}+\sigma_u\frac{\frac{\partial A_k(Y_{kt})}{\partial Y_{kt}}}{P_{ut}} = 0.\end{aligned} \tag{40}$$

We define the markup at the firm or UPC level as $\mu_k \equiv P_k \big/ \frac{\partial A_k(Y_k)}{\partial Y_k}$. Since $S_u\frac{1}{P_uY_u}=\frac{1}{\sum_k P_kY_k}$ and therefore $\sum_k S_u\frac{P_kY_k}{P_uY_u}=1$, we can rewrite equation (40) as

$$1+\left(\sigma_f-1\right)S_{ft}+\left(\sigma_u-\sigma_f\right)-\sigma_u-\left(\sigma_f-1\right)S_{ft}\frac{\sum_k\frac{\partial A_k(Y_{kt})}{\partial Y_{kt}}Y_{kt}}{\sum_k P_{kt}Y_{kt}}-\left(\sigma_u-\sigma_f\right)\frac{\sum_k\frac{\partial A_k(Y_{kt})}{\partial Y_{kt}}Y_{kt}}{\sum_k P_{kt}Y_{kt}}+\sigma_u\frac{1}{\mu_{ut}}=0.$$

Because we assume that σ_u is the same for all u produced by a firm, μ_{ut} is the only u-specific term in this expression. Hence, μ_{ut} must be constant for all u produced by firm f in time t; in other words, *markups only vary at the firm level*. Together these two results ensure the same markup across all UPCs supplied by the firm.

We can now solve for μ_{ft} by

$$\begin{aligned}1+\left(\sigma_f-1\right)S_{ft}+\left(\sigma_u-\sigma_f\right)-\sigma_u-\left(\sigma_f-1\right)S_{ft}\frac{1}{\mu_{ft}}-\left(\sigma_u-\sigma_f\right)\frac{1}{\mu_{ft}}+\sigma_u\frac{1}{\mu_{ft}} &= 0\\ \Rightarrow \mu_f=\frac{\sigma_f-\left(\sigma_f-1\right)S_{ft}}{\sigma_f-\left(\sigma_f-1\right)S_{ft}-1}.&\end{aligned}$$

A.2 Derivation of Equation (13)

Equilibrium prices P_{gt} and P_{ft}, but not P_{ut}, are functions of this N_{ft}. Recall $Y_{ut} = \varphi_{ft}^{\sigma_f - 1} \varphi_{ut}^{\sigma_u - 1} E_{gt} P_{gt}^{\sigma_f - 1} P_{ft}^{\sigma_u - \sigma_f} P_{ut}^{-\sigma_u}$. Now suppose that the number of UPCs supplied by the firm (N_{ft}) can be approximated by a continuous variable and consider the partial elasticity of Y_{ut} with respect to this measure of UPCs:

$$\begin{aligned}
\frac{\partial Y_{ut}}{\partial N_{ft}} &= \left(\sigma_f - 1\right) \frac{Y_{ut}}{P_{gt}} \frac{\partial P_{gt}}{\partial N_{ft}} + \left(\sigma_u - \sigma_f\right) \frac{Y_{kt}}{P_{ft}} \frac{\partial P_{ft}}{\partial N_{ft}}, \\
&= \left(\sigma_f - 1\right) \frac{Y_{ut}}{P_{gt}} \frac{N_{ft}}{P_{ft}} \frac{P_{ft}}{N_{ft}} \frac{\partial P_{gt}}{\partial P_{ft}} \frac{\partial P_{ft}}{\partial N_{ft}} + \left(\sigma_u - \sigma_f\right) \frac{Y_{kt}}{P_{ft}} \frac{N_{ft}}{N_{ft}} \frac{\partial P_{ft}}{\partial N_{ft}}, \\
&= \left(\sigma_f - 1\right) \left(\frac{\partial P_{gt}}{\partial P_{ft}} \frac{P_{ft}}{P_{gt}}\right) \left(\frac{\partial P_{ft}}{\partial N_{ft}} \frac{N_{ft}}{P_{ft}}\right) \frac{Y_{ut}}{N_{ft}} + \left(\sigma_u - \sigma_f\right) \left(\frac{\partial P_{ft}}{\partial N_{ft}} \frac{N_{ft}}{P_{ft}}\right) \frac{Y_{kt}}{N_{ft}},
\end{aligned}$$

where this partial elasticity holds constant the prices of existing UPCs. Using equation (3) for the firm price index and (4) for the UPC expenditure share, we have:

$$\frac{\partial P_{ft}}{\partial N_{ft}} \frac{N_{ft}}{P_{ft}} = \frac{\left(\frac{P_{N_{ft}}}{\varphi_{N_{ft}}}\right)^{1-\sigma_u}}{1-\sigma_u} P_{ft}^{\sigma_u} \frac{N_{ft}}{P_{ft}} = \frac{N_{ft}}{1-\sigma_u} \frac{\left(\frac{P_{N_{ft}}}{\varphi_{N_{ft}}}\right)^{1-\sigma_u}}{\sum_{n=1}^{N_{ft}} \left(\frac{P_{ut}}{\varphi_{ut}}\right)^{1-\sigma_u}} = \frac{N_{ft}}{1-\sigma_u} S_{N_{ft}}.$$

Recalling that expenditure shares are also the elasticities of the price index with respect to prices, we obtain:

$$\begin{aligned}
\text{Cannibalization} \equiv -\frac{\partial Y_{ut}}{\partial N_{ft}} \frac{N_{ft}}{Y_{ut}} &= - \quad \left(\sigma_f - 1\right) S_{ft} \frac{N_{ft}}{1-\sigma_u} S_{N_{ft}} - \left(\sigma_u - \sigma_f\right) \frac{N_{ft}}{1-\sigma_u} S_{N_{ft}} \\
&= \quad \left[\left(\frac{\sigma_u - \sigma_f}{\sigma_u - 1}\right) + \left(\frac{\sigma_f - 1}{\sigma_u - 1}\right) S_{ft}\right] S_{N_{ft}} N_{ft} > 0, \qquad \text{for } \sigma_u \geq \sigma_f > 1.
\end{aligned}$$

Note that this derivation is based on solely on the derivatives of the CES price index with respect to the measure of UPCs and observed prices. Since the assumption of Bertrand or Cournot competition merely affects the decomposition of observed prices into markups and marginal costs, these derivatives are the same under both forms of competition. Therefore the derivation of the cannibalization rate is the same under both forms of competition.

A.3 Cournot Quantity Competition

In our baseline specification in the main text above, we assume that firms choose prices under Bertrand competition. In this appendix, we discuss a robustness test in which we assume instead that firms choose quantities under Cournot competition. Each firm chooses the number of UPCs (N_{ft}) and their quantities (Y_{ut}) to maximize its profits:

$$\max_{N_{ft}, \{Y_{ut}\}} \Pi_{ft} = \sum_{u=\underline{u}_f}^{\underline{u}_f + N_{ft}} P_{ut} Y_{ut} - A_u\left(Y_{ut}\right) - N_{ft} h_{ut} - h_{ft}, \tag{41}$$

where we again index the UPCs supplied by the firm from the largest to the smallest in sales. From the first-order conditions for profit maximization, we obtain the equilibrium markup:

$$\mu_{ft} = \frac{\varepsilon_{ft}}{\varepsilon_{ft} - 1}, \tag{42}$$

where the firm's perceived elasticity of demand is now:

$$\varepsilon_{ft} = \frac{1}{\frac{1}{\sigma_f} - \left(\frac{1}{\sigma_f} - 1\right) S_{ft}}, \tag{43}$$

and the firm's pricing rule is:

$$P_{ut} = \mu_{ft} \frac{dA_u\left(Y_{ut}\right)}{dY_{ut}} = \mu_{ft} \left[\left(1+\delta\right) a_{ut} Y_{ut}^{\delta}\right]. \tag{44}$$

Therefore the analysis with Cournot competition in quantities is similar to that with Bertrand competition in prices, except that firms' perceived elasticities of demand (ε_{ft}) and hence their markups differ between these two cases. Since firms internalize their effects on product group aggregates, markups are again variable. Furthermore, markups only vary at the firm level for the reasons discussed in the main text above.

Our estimation procedure for $\{\sigma_U, \sigma_F, \delta\}$ remains entirely unchanged because the firm markup differences out when we take double differences between a pair of UPCs within the firm over time. Therefore our estimates of the parameters $\{\sigma_U, \sigma_F, \delta\}$ are robust to the assumption of either Bertrand or Cournot competition. Our solutions for firm and product quality $\{\varphi_U, \varphi_F, \delta\}$ are also completely unchanged, because they depend solely on observed expenditure shares and prices $\left\{S_{ut}, S_{ft}, P_{ut}\right\}$ and our estimates of the parameters $\{\varphi_U, \varphi_F, \delta\}$. Therefore whether we assume Cournot or Bertrand competition is only consequential for the decomposition of observed UPC prices (P_{ut}) into markups (μ_{ut}) and marginal costs (a_{ut}). Markups are somewhat more variable in the case of Cournot competition, which strengthens our finding of substantial departures from the atomistic monopolistically competitive markup for the largest firms within each product group.

A.4 Relative Firm Markups Under Bertrand Competition

For firms with non-negligible shares of expenditure within a product group ($0 < S_{ft} < 1$):

$$\mu_{ft} - 1 = \left(\frac{\sigma_F - (\sigma_F - 1)S_{ft}}{\sigma_F - (\sigma_F - 1)S_{ft} - 1}\right) - 1$$

For a firm with a negligible share of expenditure within a product group ($S_{ft} \approx 0$):

$$\mu_{ft} - 1 = \left(\frac{\sigma_F}{\sigma_F - 1}\right) - 1.$$

The ratio of the two markups (for σ_F finite and $S_{ft} < 1$) is:

$$\frac{1}{1 - S_{ft}}.$$

Therefore the dispersion of relative markups under Bertrand Competition depends solely on the dispersion of expenditure shares.

A.5 Relative Firm Markups Under Cournot Competition

For firms with non-negligible shares of expenditure within a product group ($0 < S_{ft} < 1$):

$$\mu_{ft} - 1 = \left(\frac{\frac{1}{\frac{1}{\sigma_F} - \left(\frac{1}{\sigma_F} - 1\right) S_{ft}}}{\frac{1}{\frac{1}{\sigma_F} - \left(\frac{1}{\sigma_F} - 1\right) S_{ft}} - 1} \right) - 1.$$

For a firm with a negligible share of expenditure within a product group ($S_{ft} \approx 0$):

$$\mu_{ft} - 1 = \left(\frac{\sigma_F}{\sigma_F - 1} \right) - 1.$$

The ratio of the two markups (for σ_F finite and $S_{ft} < 1$) is:

$$\frac{1 + S_{ft}(\sigma_F - 1)}{1 - S_{ft}}.$$

Therefore, the dispersion of relative markups under Cournot Competition is increasing in the elasticity of substitution (σ_F), and depends on both this elasticity and the dispersion of expenditure shares.