Quantity Discounts and Distortions: Scale Economies and Incomplete Pass-Through

Sharat Ganapati
Georgetown University & NBER

Colin Hottman
Federal Reserve Board

Abstract

[[very preliminary and extremely incomplete - do not distribute]]

We show the near universal prevalence of transaction-level quantity discounts in both US imports and domestic trade. In terms of aggregation, such quantity discounts account for 40% of price variation within tightly defined product categories. These quantity discounts are rooted in scale economies that are driven by transaction or search costs, and not in traditional notions of market power. Accounting for these quantity discounts imply that the pass-through of the 2018 US tariffs on imports were only 60%, with decreases in transaction sizes playing a substantial role. Furthermore, we show these estimates have implications for production function estimation.[1]

[1]Any views expressed are those of the authors and not those of the U.S. Census Bureau or the Federal Reserve Board. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product. This research was performed at a Federal Statistical Research Data Center under FSRDC Project Number 2193. (CBDRB-FY25-P2193-R12042)

1 Introduction

Markets rarely operate with uniform pricing on either inputs or outputs in both domestic and international markets (Edgeworth, 1888). The close examination of prices uncovers large variation in prices across transactions. A growing theoretical and empirical analysis as started looking into the role of buyer, seller, and relationship effects. We go beyond this and look at a more disaggregated level, the transaction, to show that a large amount of variation is rooted in quantity discounts at the individual transaction level. We further distinguish between quantity discounts and price variation due to one and two-sided market power from those that stem from quotidian transaction costs and scale economies.

Of particular importance are two central objects, the pass-through of tariffs on prices and the estimation of production functions to recover markups and productivity. We show that accounting for quantity discounts that are generated by scale economies and transaction costs significantly reduce measured pass through and markup estimates.

One particular context in which this is important is in the market for intermediate inputs, transactions between manufacturers, wholesalers, retailers, and downstream service providers. Market wedges in such markets are problematic since the core of production function estimation and growth accounting often rely on perfect competition in upstream markets. Deviations from these markets then make heavy alternative assumptions, such as forcing double marginalization, versions of Nash bargaining, or comparison with an otherwise perfectly competitive market. While all of these assumptions may make sense is a particular context, they may not be palatable in general.

Our results are straightforward. Transactions exhibit scale economies driven by the supply side. A 1% increase in a transaction's size leads to a 0.29% lower price. This is not driven by aggregate transaction volume, market level market power, or even bilateral effects. It is driven by either scale economies on either the purchasing or selling side.[2]. In terms of aggregation, while reduced form analysis may show scale economies in purchasing, and may even attribute this to monopsony power, a structural approach, adopting scale economies at the transaction level does not. Turning to one empirical application, by accounting for transaction-level scale, implies that aggregate tariff pass-through is on the order of 50%, not 100% as a reduced form analysis may imply. Similarly, production function estimation, requires assumptions on the both the flexibility of material inputs, and the invariance of quantity on prices. Correcting for this, implies differences in productivity and returns to scale.

[2]Decreasing costs on the selling side and fixed search costs on the buying side are two plausible rationales (Krolikowski and McCallum, 2025).

Empirically disentangling the market power from scale economies on the procurement side is this paper's goal. In this paper we study two contexts, the international trade and domestic trade of US firms. As any regular shopper knows, buying the multipack of household supplies is cheaper than the same good sold in a single-use package - even when the multi-pack is simply a dozen single-use items stuck together (Bornstein and Peter, 2022). The operations research literature as found copious examples of these scale economies in both procurement and sales side (Mason, 1939; Munson and Rosenblatt, 1998).

Largely speaking, aggregate economic analysis often assumes that a firm's input or outputs face a common price, common to many firms or that unit prices are such that quality is identical, or that pass-through is invariant to the supply curve, as opposed to the demand curve (De Loecker and Syverson, 2021; Ganapati, Shapiro, and Walker, 2020). There are some exceptions, for example, Atalay (2014) considers dispersion and hints at scale economies - but considers the price paid by a plant-firm-year as exogenous. Similarly, in aggregate, Davis, Grim, Haltiwanger, and Streitwieser (2013), find similar effects in electricity markets.

In macroeconomics, Bornstein and Peter (2022) find similar quantity-level discounts for consumer goods, but generalize away from any firm-level input heterogeneity. We nest their mechanism where price discounts are a function of price discrimination between downstream types. Burstein, Cravino, and Rojas (2024) find wide dispersion in input prices. They find observable characteristics of products and of buyer-seller pairs only explain a small amount of variation. We find that accounting for transactions is the missing piece.

Are these quantity discounts markups or wedges? Do they reflect either price discrimination, market power, or other wedges, such as transaction costs? or are do the reflect real costs? Bornstein and Peter (2022) assumes price discrimination. Alviarez, Fioretti, Kikkawa, and Morlacco (2023) assumes market power. Burstein et al. (2024) assumes exogenous wedges. We explore all these possibilities in US trade and some empirical support for all three. However, our results fall more in line with the distribution literature, and find that quantity discounts are driven by scale economies and fixed costs that must stem from the underlying production and demand function.

This also micro-founds the results from Gertler (2023), Amiti, Redding, and Weinstein (2019) which takes an explicitly structural approach to aggregate sectoral pricing at the country level, where pass-through is a function of explicitly modeled parameters at the firm or trading partner level.

Lastly, there is a tension in the economies of scale exhibited here and in those estimated

using production function estimation, which firm (firm-level) economies of scale closer to unity. We leave discussion on reconciliation of those facts for ongoing work.

We first decompose transaction-level prices and show that the size of transaction explains a very large portion of the dispersion in prices for domestically and internationally. Second, we show this as a largely supply-side phenomenon, where shifts in demand identify the transaction-level supply curve. Third, we show that variation in quantity discounts are driven by observable cost factors (shipping costs and within-firm transactions), as well being higher in markets with low levels of demand substitutability. We highlight both economics of scale and second-degree price discrimination, but reject differential levels of market power. Fourth, we consider aggregation and show that reduced form estimations of economies of scale and pass-through are biased, relative to our structural baseline with quantity discounts. A final section on the implications for markup estimation is forthcoming.

2 Decomposing Prices

Understanding variation in prices is one of the oldest areas of study in economics (Edgeworth, 1888). Why do prices vary? What are the implications for measurement, welfare, and efficiency? Recently, these issues have come to the forefront for two main reasons, one to understand price effects of taxes and tariffs and the second to understand the implications of market power. To better get a grip on this, we study intermediate prices at the transaction level with granular administrative and survey data. As a preliminary step, we simply decompose how much of price variation is due to various observable factors and find the dominance of one major element, simply the size of an order.

2.1 Data

We focus on two datasets. One is the universe of US imports and the other is a large representative sample of domestic transactions.

For US imports, we consider all imports that exceed the de minimis threshold and are reported to the US Customs and later processed by the US Census to add firm identifiers. We use this combined database, the Longitudinal Firm Trade Transactions Database (LFTTD), as a source for every shipment transaction entering the US via air, water, or land.[3]

[3]There are issues with using this data for pipeline based shipments.

The data considers an exporting firm (labeled as the "Manufacturer ID" from the address on the customs record, but may be a wholesaler or another intermediary) shipping products to downstream US firms. These downstream US firms are a mix of manufacturers, wholesalers, and retailers. Our unit of observation are individual transactions of an exporting firm, importing firm, 10-digit U.S. Harmonized Tariff Schedule code (HTS 10-digit), port of entry, and date of entry. For each transaction, we observe the value, shipping and freight charges, method of transport, tariffs, a standardized measure of quantity within that HTS code, and the weight of the shipment. Of note, quantities and weights are often roughly equivalent for many HTS codes, as goods are often sold by weight.

Prices are primarily computed as value/quantity, though further analysis will also compute prices using inclusive tariff and freight charges. We decompose prices for 2012, but carry out analysis using data from 2008 to 2019. In 2012, 50% of trade transactions occurs between parties that trade more than 100 times within a HS 10-digit code in a year and about 50 million transactions are reported.

For domestic trade, we consider the U.S. Census and Department of Transportation Commodity Flow Survey (CFS), focusing on 2012. This survey ask US based manufacturing and wholesale establishments about their transactions, recovering the weight, value, method of transport, and the final shipping destination. Products are classified by a 5-digit Standard Classification of Trade Goods (SCTG) code. There are three limitations to this data, compared to the import data. First, we do not observe the identity of the receiving firm, but rather only their zip-code or state, which we code as the final buyer. Second, for good that are not sold by weight, we do not see a measure of quantity - as such we drop SCTG codes for goods not sold by weight.

Lastly, this is a survey of 100,000 establishments (out of 700,000 potential establishments), covers establishments that originate 80% of domestic manufactured good shipping. Due to sampling a subset of transactions to keep reporting requirement tractable, about 5 million transactions are reported in 2012. We decompose prices for 2012, but carry out further analysis for data every 5-years from 1997 to 2017.

2.1.1 Decomposing Import Prices

To decompose prices, in each of our datasets, we residualize prices. In the international trade data, we residualize prices, quantities, and weights on the basis of exporting country and HTS 10-digit codes. For the domestic data, we residualize prices and quantities/weights on the reported 5-digit SCTG codes.[4] With this prices, we run the following

[4]We also explore the role of time variation, demeaning by HTS 10-digit and month, results are largely consistent for manufactured goods

regression to decompose prices:

Our first, and central question revolves around the role of quantity discounts. To do this, we run a series of regression with $log(price_t)$ as the dependent variable, where t reflects a transaction, or a buyer-seller-HTS 10-digit-date-port of entry tuple and $log(p)$ is the logarithm of the price, demeaned within a HTS-10 digit and Country pair.[5] The basic relationship takes the form:

$$\log p_{i \to j,t,v} = \gamma \log q_{i \to j,t,v} + \mu_{i \to j,v} + \xi_{i \to j,t,v}, \tag{1}$$

where v is a variety at the HTS 10-digit and source country level. i is selling factory. j is buying firm. t is transaction. We will use different μ fixed effects in our decomposition, with the most comprehensive fixed effect at the buyer-seller-variety level. Additionally we will allow γ_V, to flexible capture relationships that vary across different varieties.

Simply put, how much is a simple measure of order size reflected in prices? In Table 1, Panel A, we decompose our 40 million residualized prices at the transaction level. Using a uniform quantity discount across all HTS categories can account 43% of all price variation (specification 1). Following the logic of Rauch (1999), if we allow for for the relationship of log(Price) to log(Quantity) vary over different HTS 6-digit codes (of which there are approximately 5000 codes), such quantity discounts can account for 57% of all variation (specification 2).[6]

Such a decomposition implicit assumes that goods are uniform at the level of a source country and HTS product code. A stricter test would to look within the Buyer-Seller-HTS code relationship.[7] As a baseline, these highly disaggregated controls absorb 74% of the variation in prices (specification 3). However as seen in specification (6), the addition of a uniform quantity discount still can account for 37.2% of variation (16.6% directly and 20.6% through the covariance). This relationship strengthens when accounting for quantity discounts to vary at the HTS 6-digit level (Specification 4).

Even within a repeat purchases in a relationship - an extremely demanding set of nearly two million fixed effects, a very large share of price variation is explained simply by the individual transaction size.

Research on prices (cites) here, primarily focus on aggregate measures, such as buyer and seller market power, or aggregate size of purchases and sales. In Specification (6)-

[5]Samples are aligned to only include observations in all decompositions, whereby singletons are removed

[6]At the 10-digit HS level (of which there are approximately 10,000 codes, quantity discounts can account for just marginally more variation than the 6-digit level.

[7]Much of the literature has focused on buyer and seller effects. Seller-HS10 fixed effects account for 70% of the price variation in the data. Buyer-HS10 fixed effects account for 50% of the price variation.

(10), we consider two versions of these sources of variation. The principal difference is instead of considering how much is bought in a single transaction, we consider how much is bought across all transactions within a calendar year.

First we consider aggregate quantities. First, the aggregate quantity (within HTS 10-digit code), bought by a particular buyer, sold by a particular seller, and sold from a particular seller to buyer, as defined:

$$\begin{aligned}
\log\left(q_{i,v}\right) &= \log\left(\sum_{i',v'=i,v} q_{i'\to j,t,v'}\right) \\
\log\left(q_{j,v}\right) &= \log\left(\sum_{j',v'=j,v} q_{i\to j',t,v'}\right) \\
\log\left(q_{i,j,v}\right) &= \log\left(\sum_{i',j',v'=i,j,v} q_{i'\to j',t,v'}\right).
\end{aligned} \tag{2}$$

In specification 6, measures such as the total volume bought in a relationship, the volume bought by a buyer or the volume sold by a seller, collectively only explains 17.1% of the data - a far cry from the 57% explained why quantity is done at the transaction level. Furthermore, in robustness exercises in specification 7, we see that there is a minimal additional explantarory power over simply including a transaction level quantity discount.

Second we consider the role of market shares, that is what share of the quantity sold is accounted for by a buyer or what share of the quantity is purchased is accounted by a particular seller. Mechanically, we follow the definitions in equations (2):

$$\begin{aligned}
s_{i\to j,v} &= \frac{q_{i,j,v}}{q_{i,v}} \\
s_{i\leftarrow j,v} &= \frac{q_{i,j,v}}{q_{j,v}}.
\end{aligned} \tag{3}$$

In specification 8, instead of considering aggregate, we can consider share, similarly input share account for 1% of the variation alone, or 1.4% of the variation allowing for buyer HTS 10-digit and seller HTS 10-digit variation.[8] Specification 9, confirms that this variable has minimal additional explanatory power to simply having a single log quantity effect.

As a comparison to transaction-level measures, we also compare these numbers to the baselines with seller and buyer fixed effects in Specification 10 and 11. Unlike transaction level quantities, these relation-level aggregates have minimal explanatory power for the

[8]Relationship length also only has a minimal effect on explaining prices.

aggregate variation in observed prices.[9]

Table 1: Variance Decomposition of Trade Data

Panel A: US Import Trade Data

	Specification		Variance Decomposition			
	Controls	Fixed Effects	Controls	Fixed Effects	Covariance	Residual
1	$\log q$		43.6%			56.4%
2	$\log q$ x HTS6		57.4%			42.6%
3	$\log q$	Seller-Buyer		73.7%		26.3%
4	$\log q$	Seller-Buyer	16.6%	42.7%	20.6%	20.1%
5	$\log q$ x HTS6	Seller-Buyer	31.4%	32.3%	20.1%	16.4%
6	Aggregate Quantities		17.1%			82.9%
7	+ $\log q$		44.8%			55.2%
8	Aggregate Shares		0.7%			99.3%
9	+ $\log q$		43.6%			56.4%
10		Seller, Buyer			72.8%	72.8%
11	Aggregate Shares	Seller, Buyer	1.4%	72.9%	-1.3%	27.0%

Panel B: US Domestic Trade Data

	Specification		Variance Decomposition			
	Controls	Fixed Effects	Controls	Fixed Effects	Covariance	Residual
1	$\log q$		32.5%			67.5%
2	$\log q$ x SCTG5		37.6%			62.4%
4		Seller-Buyer		73.7%		26.4%
5	$\log q$	Seller-Buyer	18.6%	47.3%	12.0%	22.1%
6	$\log q$ x SCTG5	Seller-Buyer	20.4%	45.6%	12.4%	21.6%

Notes: Panel A decomposes transaction-level price variation after demeaning log(prices) by country-origin and product code (HTS 10-digit) fixed effects. Sellers are designated at the manufacturer-id and product code level. Buyers are designated at the domestic firmid and product code level. Panel B decomposes transaction-level price variation after demeaning log(price) by product code (SCTG 5-digit). Only products where quantities are regularly denominated in weight are included. Sellers are designated at the establishment and product code level. Buyers are designated at the destination state and product code level. For consistency, the sample (both within Panel A and within Panel B) is fixed to remain constant over the sample period. See the text for full details and specification.

Wrapping up, accounting for the size of an order, within a tightly defined product group can alone account for nearly 60% of price variation in international trade, 50% of the within-variation after accounting for every buyer-seller-product relation. While relationship controls (buyer-seller factors) can explain some price variation, a significant amount is still explained by the size of a individual transaction. Does the same relation-

[9] In futher robustness checks, we show that the price variation imbued in other relationship level outcomes, such as the length of relationship and intra-firm trade have even lower explanatory power.

ship hold within domestic trade?

2.1.2 Decomposing Domestic Trade Prices

We conduct a similar exercise with the US domestic Commodity Flow Survey (CFS). While limited in certain aspects, particularly since it is a highly limited sample from the buyer's perspective.[10]. We broadly find similar trends to the international trade import data.[11]

Data on quantities is only restricted to material sold by weight, as a quantities are not recorded (as in international trade). We subset analysis only to goods sold traditionally by weight.[12] Unlike US import data, data is not collected on the identity of the buying firm, only the identity of the seller is know. We only know the address of the firm receiving shipment. We will simply use the ZIP code of the recipient firm as a proxy for the buyer. Product codes are substantially more aggregated. The 5-digit SCTG codes are largely comparable to a 4-digit HTS import code or 4-digit HS trade code.

In Panel B of Table 1, Row (1) replicates the exercise to decompose the variance of prices on the size of a transaction (all de-meaned and residualized at the SCTG 5-digit level). Transactions are identified with a seller, product code, shipment date and buying zip code tuple. [13]

Specification (1) finds that 32.5% of all price variation is due to a log-linear quantity relationship. Allowing this relationship to vary across SCTG 5-digit products, increases this to 37.6% of all variation. The share of prices explained by quantity are indeed a bit lower than those in the international trade data, however that data has better measures of products categories and direct measures of quantities. Regardless of the data generating process, within tightly defined product groups, a simple log(q) explains between 30-40% of all price variation.

Specifications (3)-(5) attempt to see if this relationship is purely reflective of buyer-seller pair, or if there is any explanatory power due to transaction-level quantities. Broadly, as with international trade, the results are similar.

We further replicate this analysis using the 2012 public use data in a forthcoming appendix.

[10] This is important since XXX and XXX assume away price variation in input data

[11] This analysis is replicable on the public use data, that aggregate regions and SCTG codes to a higher level of aggregation

[12] Interestingly loosening this restriction does not significantly change our results

[13] Mechanically, X% of transactions in a given year are 'one-off' between a unique seller, product code, and buying zip code, but they only account fo XX% of value.

Alternative Samples: Export Trade Data As an alternative (as the structural approach in Gertler (2023)), we can use export data. One downside of this data is that is does not have the identity of the buyer, but rather just their location. We find nearly the level of explanatory power relationship.

Discussion At this point, regressions and decompositions are purely descriptive. They are there to attempt to parsimoniously describe an equilibrium relationship.

Simply knowing how much you bought, rather than who you the buyer is, is a better predictor of a transaction's price. Futher the explanatory power of the transaction size remains relavant, even with a massive number of relationship fixed effects. While a vast literature exists on seller and buyer effects, we are going to focus on quantity effects. Within a product code, over 50% of prices can be explained by log(q), the quantity bought.

Instead of looking of the "effects of causes", we are aiming to understand the "causes of effects", as such we now turn to a causal anaylsis. Can we add economics to understand that an exogenous increase in order size leads to a decrease in prices - as opposed to a story where we are simply recovering a demand schedule. We first turn to making this relationship causal, is this about demand, supply, or the interaction between the two? In particular, we focus on the supply side, noting that demand is already typically downward sloping.

Our next set of analysis attempts to recover the pricing schedule faced by buyers. First in a parsimonious specification, leveraging time-series variation in demand schedule

3 Scale Economies and Quantity Discounts in Purchases

We find a large driver of price variation at the transaction level is largely linked to transaction size, even after controlling for buyers and sellers within product groups. The next goal is to understand a causal link. We are going to focus on isolating the supply side from the demand side. On the supply side, do higher quantities lead to lower prices. Or on the demand size, do lower prices lead to higher quantities.

We start at a fundamental building block of firm to firm linkages - a single transaction and aggregate upwards. First to a bilateral relationship, then to aggregate industry-level relationships, both at the intra-national and international levels.

As in the price decompositions, the basic relationship takes the form:

$$\log p_{i\to j,t,v} = \gamma \log q_{i\to j,t,v} + \mu_{i\to j,v} + \xi_{i\to j,t,v}, \tag{4}$$

where v is a variety. i is selling factory. j is buying firm. t is transaction. At the baseline there is a fixed effect at the buyer-seller-variety level. Identification comes off different prices at different quantities within a relationship. In a forthcoming appendix, we micro-found downstream demand shocks affecting the optimal inventory stocking of a downstream firm, extending the framework of Arrow, Harris, and Marschak (1951). This leads to firms ordering different levels of product over the course of a year.[14]

This elasticity γ, or inverse supply elasticity, governs how quickly prices fall as demand increases - within a buyer-seller relationship.

This equation is perhaps one of the most commonly estimated equation in economics - what is the supply curve? This strategy is in the spirit of the work reviewed in Hillberry and Hummels (2013), but differs in a critical element - we consider transaction and abstract away from any measures of aggregates of transactions or direct measures of classic market power. At this point, we are looking at an equilibrium relationship, before considering their link to traditional models of either aggregate trade relationships or pre-specified models of firm conduct.

Previewing results, baseline estimation finds values of γ around -.28. A doubling of a purchase realizes in an approximately 28% decrease in the purchase price within a transaction. Converting this estimated relationship to a causal one has two common threats. First, unobserved product characteristics, even within a buyer-seller-10-digit HTS code triad (or relationship). Second, are we mistakenly absorbing potential changes in the supply curve - as opposed to helpful shifts in the demand curve. This would imply we are mistakenly simply recovering a traditional demand curve. We outline a strategy to control or instrument for these major issues.

3.1 Identification

Unobserved Product Characteristics Are we mis-measuring quality within even the highly dis-aggregated HS-10 categories and a trading pair of downstream buyers and factories. An example being we may buy 10,000 iPhone 16s at $500 and 1,000 iPhone 17s at $1000 and both under HS code "8517.14.0050". If goods are sold in efficiency units, we need to find a shift for quantity, uncorrelated with quality, but not correlated with ξ. Note this is the exact opposite of the problem that we traditionally worry about in industrial organization demand estimation, where price is positively correlated with unobserved quality. Mechanically if quantity purchased varies inversely with quality, we may find a

[14]This model also produces variation in the time between orders, which in theory is a testable implication. This also ties in with work linking inventories with international trade, as in Alessandria, Kaboski, and Midrigan (2011).

mechanical negative relationship.

This would be consistent with some degree of third degree price discrimination. It also would be consistent with a form of the Alchain-Allen "shipping the good apples out" effect (Hummels and Skiba, 2004). However either effect would be a threat to identifying a form of the supply curve. (This exists for even relatively homogeneous products, such as concrete and gasoline). Essentially we need a time-varying demand shock, either at the buying firm level, or even better at the buying-firm HTS10-digit level or buying firm-sourcing country level.[15]

Instrumental Strategy This comes down to $Var(q, \xi)$. Do you buy more or fewer units of higher quality products? If your demand is inelastic and products are sold in efficiency units, is this correlation is positive?

We aim to find an instrument for $q_{i \to j,t,v}$, with our high degree of fixed effects, we need a very fine grained instrument, one that varies essentially at the daily or weekly transaction level. At this point it helps to put further structure on demand. If we assume that within a HS-10 code and buyer, goods are either gross complements or substitutes, then changes in prices and quantities bought can act as instruments for the quantity from a particular transaction. (Write out this demand system)

We construct an IV as follows;

$$IV_{i \to j,t,v} = \log \sum_{i \neq i', t' \in (month_t)} q_{i' \to j,t',v} \tag{5}$$

We construct this instrument three different ways. As the mainline specification, we look at a buyer and HTS 10-digit code and consider within the month, how much they bought of that variety from the same supplier, leaving out the current shipment. Then as a robustness, we consider a two variants. First we look how much a buyer bought of a variety within a month, but leaving all purchases from the same supplier out. Second, we consider how much a buyer bought in aggregate of a product, but

However, there remains an issue, are there higher order network effects, causing a buyer to switch from a particular supplier, causing an order size change? To account for that, we to a direct supply-side control. Furthermore, while this does largely control for unobserved product quality, it may still not fully control for supply side shocks.

[15] We did experiment with using exchange rate variation, but exchange rate variation at the weekly or monthly level did a poor explanatory job - see papers on invoicing and lead time

Table 2: Recovering Quantity Discounts: Import Data

	(1)	(2)	(3)	(4)	(5)
			log p		
log q	-0.454	-0.282			-0.274
	(0.00254)	(0.00253)			(0.00275)
log q x HTS-6			-0.364	-0.293	
			[0.141]	[0.130]	
r2	0.437	0.797	D	D	0.811
Within r^2	0.437	0.24	D	D	0.232
Fixed Effects					
Relationship		X		X	X
Manuf-time-prod					X
Demeaned					
Seasonality (HTS-month)	X	X	X	X	X
Variety (HTS-county)	X	X	X	X	X

	(6)	(7)	(8)	(9)	(10)
			log p		
log q	-0.258	-0.254	-0.278	-0.282	-0.305
	(0.0000928)	(0.0000952)	(0.0000364)	(0.0000357)	(0.000411)
log(Unit Weight)					0.476
					(0.00243)
r^2	0.772	0.801	0.211	0.210	0.456
Within r^2	0.217	0.213			
Fixed Effects					
Relationship	X	X	X	X	X
Manuf-time-prod		X	X	X	X
Demeaned					
Seasonality (HTS-month)	X	X	X	X	X
Variety (HTS-county)	X	X	X	X	X
First Stage F-Stat			1900000	670000	11000
Instruments			Main	Variations	Main

Notes: Round parenthesis represent standard deviations. Square brackets represent the variance across HTS 6-digit codes for estimates. The mean estimate is displayed over it. Demeaning refer to process of regularizing all variables by country-origin and product code (HTS 10-digit) fixed effects, as well as month and product code fixed effects. Sellers are designated at the manufacturer-id and product code level. Buyers are designated at the domestic firmid and product code level. (Table 1 does not include seasonality effects). See the text for full details and specification. D denotes variables not yet disclosed from census. Standard errors are clustered by the relationship.

Supply Shocks Is the identifying variation coming off a shift in the supply curve, rather than a shift in demand? This is a threat since we could be solely identifying a demand curve if there are no demand shifts, or we are identifying neither as we are simply plotting all possible equilibriums in a 2-dimensional space. While the instrument proposed implicitly only considers variation in the quantity supplied due to demand shifts, we can directly control for control for supply shocks. These trading relationships are long-lived, in practice we often see hundreds or thousands of shipments between trading pairs.

We control directly for $\gamma_{i,v,t\in(t-1,t+1)}$, so and demand shock must be very high response. We do this at the month level. One downside of this fixed effect is its stringency, leading to a moderate drop in the sample size. [16]

There still remains a threat that there are extremely high frequency supply side shocks (at the weekly level), that are not industry specific, but specific to a single supplier, and that do not affect past or future shipments within that month. We leave that discussion for future work.[17]

3.2 Transaction-level Supply: Import Regression Results

We display the results of these regressions in Table 2. Column (1) display the results of a simple regression of demeaned quantities on demeaned prices. Absent either structure or a method of considering the root of the variation, this could reflect supply curves, demand curves, or some other complex interaction. A doubling of the transaction quantity is associated with a 45% reduction in prices. Column (3) repeats this exercise, but allows for variation with a HTS 6-digit code. On average, the average product exhibits a relationship where a doubling of a transaction quantity is associated with a 36% decreasing in prices. This negative relationship exhibits a variation of 0.141 around this mean result. Approximately 95% of HTS 6-digit codes exhibit a negative relationship.

To make progress, column (2) exploits the predominately repeated nature of the underlying data and controls for a relationship. That is a buyer-seller-product (at the 10-digit HTS code level) fixed effect. This can absorb a large degree of unobserved product variation, as well as buyer, seller, or bilateral market power indicators. Within a relationship, after demeaning variety and seasonality effects, a doubling in the order size is correlated

[16] We can more broadly control for time varying country-level characteristics, such as $\gamma_{c,v,t}$. As a robustness exercise, this does increase the sample size, but with no significant effect on the estimated coefficients. Similarly, within a year, we can control for 'relationship' length and again find no difference in results. Some papers subset this to looking at arms-length relationship. We find no differential effect there. We further explore this in the next mechanism section.

[17] A full analysis of such high frequency shocks would need a full model of inventories, demand and shipping lead times.

with a 28.2% decrease in reported transaction prices. Column (4) allows for γ_v to vary across aggregated HTS codes. Results are largely unchanged, with the mean effect across products indicated a doubling of order sizs is associated with a 29.3% in prices. Finally to absorb time variation in supply, column (5) adds supplier-time (at the monthly level) fixed effects. Results are largely unchanged, with a doubling in the order size associated with a 27.4% decrease in prices.

While partially correcting for supply shocks in specification (5), unobserved quality still acts as a threat to the causal identification of a within-relationship supply side. The bottom panel in Table 2 shows the results of our instrumenatal variable strategy. Of note, as the instrumental variable strategy does not always provide an instrument, our sample does change. Specification (6) replicates specification (2), but with a consistent sample and specification (7) replicates (5). Broadly speaking, the point estimates are slightly attenuated, but consistent. A doubling of the order size is reflected in a 25.4% increase in price.

Column (8) implements our main instrument from equation 5. The first stage demonstrates a strong effect. The final point effect is largely unchanged, estimating the elasticity γ of -0.278. A robustness exercise replacing our baseline instrument that excludes transactions with thee same supplier to only look at other suppliers finds similar results in column (9).[18] There is a slightly lower first stage F-value, but the point estimate is largely unchanged. Finally, as a direct control for quality, we consider only products that are not sold by weight (that is products where the quantity sold is not within 10% of the reported shipping weight). As weight may be an indication of quality (higher weight may reflect higher quality materials or in other cases, higher weight may reflect lower quality materials), column (10) shows this has a minimal effect on our point estimates.

3.3 Supply Estimation: Domestic Data

We repeat a version of the supply isolation using the domestic shipment data from the CFS. The available data makes it hard to directly control for supply shocks, but an aspect of the data collection process works in our favor. In particular the government questionnaire recovers a sample of shipments, not throughout the year, but within a tightly defined window (typically a week or so). This mechanically controls for supply-time fixed effects. As such all supply shocks (after controlling for the relationship), must differentially affect some, but not all other customers. As such, our primary identification threat are unobserved qualities across shipments.

[18]The idea is that a shift in another supplier should not affect the focal supplier firm.

As before, we create an instrumental variable that is correlated with a demand shock, but uncorrelated with the current supplier (for which we have implcitly a supplier-time control). Following Equation 5, we create three instruments. The first looks at a destination state, excluding that order, how much of that product is bought from that firm. This naturally is at roughly the month level, as orders from firms are clustered due to sampling. The second, considers how much is bought by a destination state within a month-product, excluding the focal transaction. The last, does the same, but excludes all transactions from the original focal firm.

Table 3, produces a version of Table 2 for our domestic data in 2012. Column (1) illustrates in the simply demeaned data a correlation, where γ is equivalent to -30.4%; broadly in line with the domestic data. However, looking within a buyer-seller-product relationship in column (2), the same γ is -0.23. Columns (3) and (4) repeat this exercise, but allow γ_v to vary across each of the 5-digit SCTG products. Mean estimates of γ_v are -0.295 and -.206, with relatively tight variances acros different products.

Turning to our instrumentation strategy to control for unobserved product quality, as well as the limited possibility of supplier shocks, columns (5)-(6) display the results. Column (5) replicates column (2), but with the same sample as observations with out first instrument. Column (6)-(7) show that all three instrumentation strategies yeild strong predictive first stages and have γ_v between -0.21 and -0.26.

Broadly, within-relationship inverse supply elasticities are around -0.28 in the trade data and -0.24 in the domestic trade data. However the good traded are different and product level comparisons may yield even more consistent estimates.

With this limited regression approach still potentially leading to concerns about endogeneity, we turn to another approach, which adopts a widely used strict supply and demand side framework to recover within-relationship supply elasticities.

3.4 Alternative Estimation Techniques: Feenstra 94

One downside of this reduced form approach is that it doesn't account for the shape of demand (such as the elasticity of substitution), but rather instruments for variation in demand while simultaneously controlling for supply shocks. An alternative approach is to structurally estimate supply and demand within a relationship over time. Following the trade literature, we can estimate both the within-relationship (inverse) supply elasticity alongside the demand elasticity using panel, which we collect at the monthly level for relationships. This is a simplification from our purely transaction-level approach, but as the estimation require a continuous panel, we compromise at this level of aggregation.

Table 3: Recovering Quantity Discounts: Domestic Data

	(1)	(2)	(3)	(4)
	log p			
log q	-0.304	-0.23		
	(0.00146)	(0.00174)		
log q x SCTG			-0.295	-0.206
			[0.136]	[0.0246]
r^2	0.325	0.779	0.376	0.784
Within r^2	0.325	0.159	D	D
Fixed Effects				
Relationship		X		X

	(5)	(6)	(7)	(8)
	log p			
log q	-0.22	-0.209	-0.256	-0.261
	(0.00197)	(0.00170)	(0.00423)	(0.00508)
r^2	0.788	0.155	0.151	0.15
Within r^2	0.155			
Fixed Effects				
Relationship	X	X	X	X
First Stage F-Stat		74000	69000	11000
Instruments		IV1	IV2	IV3

Notes: Round parenthesis represent standard deviations. Square brackets represent the variance across SCTG 5-digit codes for estimates. The mean estimate is displayed over it. Demeaning refer to process of regularizing all variables by product code (HTS 10-digit) fixed effects. Sellers are designated at the establishment and product code level. Buyers are designated at the domestic destination state and product code level. See the text for full details and specification. D denotes variables not yet disclosed from census. Standard errors are clustered by the buyer-seller pair.

Similarly, we also run our analysis at the 6-digit HS code level.

As an alternative, we can put structure on demand (say CES), we can jointly estimate supply and demand. (Following Feenstra (1994), and in an application similar to Hottman, Redding, and Weinstein (2016)), as opposed to the solely supply side exercise with an instrament for demand. There is one downside to this, all demand and supply functions much be globally pre-specified. We will assume constant elasticity of demand (CES) and a constant elasticity of supply (globally as opposed to locally).

As for identification, fundamentally, there is a timing assumption, shocks in demand between time periods and between trading pairs are orthogonal to shocks on the sup-

ply side. This assumption does need a parameterization of both supply and demand to recover residuals.

Traditionally, this analysis is run at the origin-country, destination-country, and product category level (extensions include Broda and Weinstein (2010) and Soderbery (2015)).

Formally, assume the transaction-level demand is CES. Taking logs, taking the time difference and differencing relative to another buyer-seller pair r in the same sector s gives

$$\Delta^{r,t} \ln(q_{vt}) = (-\sigma^s)\Delta^{r,t} \ln(p_{vt}) + \nu_{vt}, \tag{6}$$

where $\Delta^{r,t}$ refers to the double difference.

Next, we assume the following transaction-level pricing equation holds in double-differenced form:

$$\Delta^{r,t} \ln p_{vt} = \omega^s \Delta^{r,t} \ln(q_{vt}) + \kappa_{vt}, \tag{7}$$

We assume that the following orthogonality condition holds for each buyer-seller pair:

$$G(\beta_s) = \mathbb{E}_{\mathbb{T}} \left[\nu_{vt}\kappa_{vt}(\beta_s)\right] = 0 \tag{8}$$

where $\mathbb{E}_{\mathbb{T}}$ is the time series expectation and $\beta_s = \begin{pmatrix} \sigma^s \\ \omega^s \end{pmatrix}$.

Re-writing this orthogonality assumption gives:

$$\mathbb{E}_{\mathbb{T}} \left[(\Delta^{r,t} \ln p_{vt})^2\right] = \mathbb{E}_{\mathbb{T}} \left[(\frac{\omega^s}{\sigma^s})(\Delta^{r,t} \ln q_{vt})^2 + (\frac{\omega^s\sigma^s - 1}{\sigma^s})\Delta^{r,t} \ln q_{vt}\Delta^{r,t} \ln p_{vt} + \frac{1}{\sigma^s}\nu_{vt}\kappa_{vt}\right], \tag{9}$$

which simplifies to:

$$\mathbb{E}_{\mathbb{T}} \left[(\Delta^{r,t} \ln p_{vt})^2\right] = (\frac{\omega^s}{\sigma^s})\mathbb{E}_{\mathbb{T}} \left[(\Delta^{r,t} \ln q_{vt})^2\right] + (\frac{\omega^s\sigma^s - 1}{\sigma^s})\mathbb{E}_{\mathbb{T}} \left[\Delta^{r,t} \ln q_{vt}\Delta^{r,t} \ln p_{vt}\right]. \tag{10}$$

Implementation Implementation of the simultaneous identification of supply and demand is relatively demanding. It requires a continuous time series for multiple buyers and sellers over time.[19] Furthermore the identification of such demand and supply curves are not directly applicable to the idiosyncratic transaction level data. To adapt the data, we make for two adaptations. First, instead of considering products at the root 10-digit level. We aggregate to the 6-digit level. Second, instead of using the raw transactions at

[19]This technique is infeasible with the domestic data, as data is collected only in a single time period for any origin.

the daily level, we aggregate to the monthly level. [20]. Overall, we are left of a time series of buyers and sellers from 2008 to 2016, where the relationship is active for 12 continuous time periods (months in this application). We winsorize price, sales, and quantity changes at the 1% and 99% levels.

There are also a few choices that need to be made for GMM estimation. First is the initial search space. We search over values σ and ω in the space starting by -2 to 2 for ω and 1 to 12 for σ. Second is in the weighting of observations. We largly follow Broda and Weinstein (2006) and weight observations by a function of the count of observations within a trading relationship.

Results Operationally, this technique is more demanding than that in Section 3.2, furthermore it requires a grid search to account for multiple solutions. Following the literature, we do primaily restrict the solution to those with downward sloping demand; but are flexible in terms of supply.

As noted in the literature, there are multiple possible solutions, due to the fact that the solution lies at the intersection of two hyperbolas. We find this common. Typically we find there are two solutions, often one of the solutions is small and with ω near zero and the other solution is broadly falls in the line of our estimates, were $\omega \approx \gamma_v \approx -0.3$.

Table 4: Combined Supply and Demand Estimation

	Supply Elasticity (ω)			**Demand Elasticity (σ)**		Correlation
Spec	Median	Mean	Variance	$\frac{1}{mean(\sigma^{-1})}$	$e^{mean(\ln\sigma)}$	(ω,σ^{-1})
1	-0.233	-0.365	0.205	3.425	7.794	0.587
2	-0.618	-0.496	0.183	6.494	11.510	0.465
3	-0.222	-0.357	0.208	3.497	9.976	0.532
4	-0.007	-0.073	0.163	1.751	3.076	0.053

Notes: See text for details.

Table 4 shows our results. Fundamentally, ω is similar to the γ estimated, but with a different strategy. Here instead of relying on real-world instrumentation and fixed effects, identification instead relies on the orthogonality of the variance of supply and demand. We consider this.

We first choose the solution with $\sigma < 0$ and with the smallest inverse supply elasticity. In this main specification (column 1), finds that a median elasticity of -.233 and a median

[20] A forthcoming appendix shows that such an adaptation is not problematic under relatively mild conditions, and may realize in an slight underestimate of the supply elasticity

elasticity of -.365, with a variance of 0.205 across HTS codes. These estimates are broadly in line with our standard identification strategy and illustrates the robustness of both the instrumentation strategy and well as the consistent identification of demand.

In terms of the demand side, there is a wide right tail in σ, with some point estimates trailing off towards infinity. Some researcher use ad-hoc trimming of the underlying data to get that under control. Instead we report the inverse of the mean of the inverse. This transformation (and the related logramthic and exponential) transformation, finds a mean demand elasticity of approximately of $\sigma = 3.4$, which is broadly in line with the literature.[21] Furthermore, the supply elasticity ω has a correlation of 0.47 with that we recover using the reduced form methodology (the parameter γ).

Specification (2), we do restrict σ to being positive, and has results broadly consistent, however a $\sigma < 0$ often has unclear economic properties. Specification (3) drops products with σ that do not converge, results are still broadly consistent. As an alterative, we use the maximum estimated ω in specification 4. However, the corresponding σ are also extremely small, and estimates are often extremely imprecise.

Overall, this is broadly consistent with the supply elasticities estimated in the more reduced-form exercise Section 3.2, as well as firm-demand estimates of σ from Hottman et al. (2016).[22]

3.5 Discussion: Transaction Level Supply

The more you buy, the cheaper it is. This exists for all large-scale material purchases in the United States, regardless if they are sourced domestically or internationally. While there is some variation across good. The vast majority of HS-lines in imports demonstrate this characteristic. Similarly, most SCTG-codes accounting domestic physically shipments from wholesalers and manufacturers also demonstrate this characteristic. This exists both in the a within-specification using ordinary least squares, as well as when controlling for supply shocks and with an instrumental variable strategy based on complementary or substitutability across products. This scale economy on supply exists at the transaction level and matters for firms making purchase choices.[23]

We first turn to what is driving this scale economy in purchases, before considering aggregate implications. Additionally, we build off the initial findings that there is variation across different product types and consider

[21] This is similar to the median, but which we have not disclosed from the Census.

[22] It also lines up broadly with Broda and Weinstein (2006) and others at the level of substitution between countries

[23] A separate but linked literature looks at capacity utilization.

4 The Mechanisms Behind Quantity Discounts

What is the underlying reason behind these quantity discounts or scale economies in purchasing inputs? We approach this problem from the perspective of three broad classes of explanations. The first are - does this reflect some bilateral relationships; perhaps bilateral market power as in Alviarez et al. (2023)? Or something about the longevity of relationships and contractual details as in Antras and Helpman (2004)? Second, does this reflect one-sided effects, such as classic buyer or seller fixed effects; such as monopoly or monopsony? Lastly, are these quantity discounts varying systematically across markets? Or do they primarily reflect second degree price discrimination?

Furthermore, our end goal is to determine, or at least eliminate potential causes for quantity discounts. Are they reflective of underling costs, various forms of market power, or some interaction between the two?

This type of quantity discount or second-degree price discrimination can occur for two broad classes of reasons. The first class of reasons is dependent on cost structures and the second class is reliant on market power to extract consumer surplus from heterogeneous customers. [24] Fundamentally the difficulty in separating the two is in the absence of cost data at the transaction level.

In particular, we consider the transaction-level quantity discount (as opposed to an aggregate purchase quantity). This is justified in a variety of ways, largely falling along two different lines. The first is due to due to differences in underlying costs and economies of scale. The second is largely a form of second degree price discrimination from a seller with a degree of market power. At first, we will not take a stance on this reasoning, revisiting this at the end of this section. Rather, we simple aim to measure if this quantity discount exists and is quantifiable.[25]

These quantity discounts are a fundamental issue in competition and welfare economics. Fundamentally, regulators care (Robinson Patman Act in the U.S. as in Ostlund (1939)). For example, quantity discounts that do not reflect market power are legal.[26]

[24] Buyer-paid search costs can produce the second type of result as in Marshall (2020) looking at restaurant suppliers.

[25] We focus on the trade data here. When possible results are largely replicated in the domestic CFS data, but as that data is a survey - not all analysis can be fully replicated.

[26] Of particular concern is not upstream market power, but rather downstream market power. though the monoposony literature has changed some perspective in the united states

4.1 Bilateral Measures

What role to does the relationship between the buyer play on the quantity discount? There is a large literature touching on the role played on price levels due to such effects, such as bilateral monopolies, vertical integration, and monoposony. But broadly speaking, most empirical work generalizes from economies of scale at the transaction level, and simply focuses on the realized prices between a particular buyer and seller. As we later show in Section 5, understanding these effects at the transaction level has implications for aggregate prices. We focus on two four types of bilateral indicator. First, vertical integration. Second, relative input and output market shares. Third, economies of scale within a relationship. Fourth, measures of relationship length.

Vertical Integration A large literature considers the role of vertical integration within a firm on prices, primarily at the aggregate level, eventually considering the effect on both within-firm prices, as well as aggregate market effects. We as a more basic level, do we see any effect of differential transaction-level quantity discounts between within and between-firm transactions. Using the the US Custom definition of related parties as our indicator[27], in the top left regression (1) in Table 5, we report on the differential effect of a related parties indicator on the quantity discount (calculated within a HTS 6-digit category). To isolated the quantity discount effect from a price level effect, we include buyer-seller-product (HTS 10-digit) controls. Broadly speaking, the related party indicator indicates a 0.6% lower quantity discount, relative to a baseline of 30%. While statistically significant, the economic effect is relatively negligible and is very low in terms of explanatory power, with a negligible added r^2

Taken at face value, if we assume that transaction prices within a firm reflect underlying costs, the lack of a differential quantity discount implies that quantity discounts at the transaction level reflect differences in costs, stemming from either a notion of transaction level fixed costs or economies of scale.

Of note, the particular laws governing how valuations of products within a relationship may simply mean that the related party flag itself is not an indication of differential application of market power. Transaction value can be determined by two broad methods, one is pricing is determined using a method of appraisal and the other is if pricing is consistent with pricing in the industry and to non-related buyers. One permissible method is using actual cost measures plus a fixed profit share computed at the firm-level. If prices are determined in this manner, the quantity discount for intra-firm trade reflect

[27]

true cost differences, as they are absorbed by the fixed effects. However, if prices are determined using a reference system, and that reference system implicitly includes some constant measure of market power that is differential between transaction-sizes, vertical integration prices differences may not reflect true differences in market power due to legal constraints on reporting.[28]

Due to this caveat, we turn turn to measures of bilateral market power using computed measures of market power[29]

Bilateral Market Shares Theoretically, if a buyer accounts for a large market share of a seller, or a seller accounts for a large market share of a buyer, this may reflect bilateral market power, as there are differential outside options for the buyer and seller. There are many ways to implement this, often using the Nash-in-Nash bargaining framework (see healthcare studies, but also Moralcco et al in trade). Supplier dominance is predicted to lead to higher prices and buyer dominance is predicted to lead to lower prices.

In our context, do these effects also show up in transaction-level quantity discounts? Do sellers give lower quantity discounts when they have bilateral market power - as they can now prevent dependent larger buyers from bargaining? Using the buyer and seller market shares from equation 3 in terms of the quantity sold[30], $s_{i\rightarrow j,v}$ and $s_{i\leftarrow j,v}$, regressions (2) and (3) implement this test in Table 5.

As with the related party indicator, we do find a statistically significant effect, with low economic significance, with explanatory power. Within relationship buyer market power is linked to a slightly larger quantity discount. Going from zero to 100% market share for a buyer is reflected in a 1.5% bigger quantity discount (off a base of an 30% quantity discount). For seller market power, that is considering sellers who dominate a buyer's imports, going from 0 to 100% market share is associated with a 2% quantity discount. But as before, there is an extremely small within variation for both variables.

What if, instead of market power, costs reflected fixed costs, not at the transaction level, but at the relationship level.

[28]Of note though is that at the aggregate level, conditioning for order size, a seller received a lower price for within-firm transactions and a buyer pays higher prices for non-within-firm transactions, suggesting the lack of enforcement or even measurement issues.

[29]We find that dropping related party transactions makes a minimal difference in all the regression in this paper. For completeness, we keep these observations in.

[30]Moralcco et al use seller revenue share and buyer quantity share; the risk is in such models, you may mechanically get a positive effect on revenue share and a negative effect on quantity share with buyer and seller fixed effects as $\log(p) = \log(rev) - \log(qty)$ and $\log(rev_share) = \log(rev) - FE_{seller}$ with $\log(qty_share) = \log(q) - FE_{buyer}$

Relationship Discounts Are within-relationship quantity discounts bigger or smaller for big versus small relationships? Theoretically, a big buyer may get bigger quantity discounts, reflecting input shares. On the other hand, big buyers may see smaller quantity discounts, as reflecting the fact that switching may be more costly. Similarly, what about long-lived versus shorter and newer relationships? Regressions (4) and (5), illustrate that big relationships and longer relationships exhibit smaller quantity discounts within the relationship. However, the measured effects are relatively small and explain little of the variation in prices, as opposed to the aggregate average quantity discount.

4.2 Unilateral Measures

Bilateral measures do little to explain quantity discount schedules within relationships. But a broad unilateral literature may have explanatory power. Why are prices decreasing in order size? One classic story is monopsony, in fact in the labor literature this is often considered a primary driver of wage markdowns (Manning, 2013; Sokolova and Sorensen, 2021) (Manning 2003. Larger buying firms can exert market power to pay less.

Seller Size or Market Power We consider seller market power in three different ways. One is simply looking at quantity discounts for big versus small sellers. The second is looking at market shares of sellers. The third is looking at the squares of market shares (the HHI weights). Broadly speaking, one source of second degree price discrimination quantity discounts stems from seller market power, were sellers can discriminate against price insensitive buyers who may buy smaller quantities. Regressions (6), (8), and (10) in Table 5 shows our estimated relationship. Both measures of market share exhibit a small, but economically insignificant effect. The size of a seller shows that big sellers have smaller quantity discounts, which works against the conventional story, as bigger sellers should better at price discrimination. However, this effect is still relatively small in magnitude and explanatory power.

Buyer Size or Market Power As a counterpart to seller market power, Table 5 Regressions (7), (9) and (11) show the corresponding relationships. Both measures of market share exhibit a small, but economically insignificant effect. Larger buyers again face smaller quantity discounts, but with a small explanatory effect. The fact that all three of the seller's size, buyer's size, and the relationship size all exhibit decreased quantity discounts show there may be a relationship level for large buyers and sellers that may substitute for transaction level effects.

Essentially, there is a very small of bilateral or unilateral effects on the magnitude of the quantity discount. There is variation in quantity discounts, but as seen in Table XXX, this variation seem to reflect simple differences across product types, not within product types. With that in mind - then what does that leave? One potential set of characteristics are at the market level. Empirically, quantity discounts and variation in quantity discounts must be driven by variation that appear at the market level. What characteristics of

4.3 Market Level Characteristics

Are market-level quantity discounts driven by markets that systematically vary in observable characteristics, in the number of buyers, sellers, or pairs? Or are they driven by some set of characteristics about the supply or demand sides? As in the old supply-conduct-structure paradigm, reading into this analysis should not be viewed in a purely causal lens. We will run cross-sectional regressions across products to try to simply understand if there are characteristics than help discipline theories. As before, the the observed quantity discount is an equilibrium object, with entry and supply choices determined by underlying cost and preference function.

We run correlational regressions of the form:

$$\gamma_v = \alpha Y_v + \epsilon_v, \tag{11}$$

where γ is the recovered quantity discount from Table X, regression X for variety v which indexes a particular HTS 6-digit product code. If we condition on market entry and firm size as exogenous, the first five columns of Table 7 considers the effects of markets with different seller HHI, buyer HHI, number of US buyers, number of foreign firms selling to the US, and number of Foreign sellers-US buyer pairs. Broadly speaking, there is no measurable effect for any of these variables. Conventional measures of entry, buyer-seller thickness, or market power have any observable correlation.

In Regressions (6)-(10) in Table 7, we consider attributes more reflective of underlying processes of supply and demand. Column (6) shows that markets with lots of transactions have smaller quantity discounts. Column (7) shows that large markets with large transaction values, similarly have smaller quantity discounts. Column (8) incorporates our estimates of a supply elasticity from Section 3.2. Column (10) adds in estimates of demand elasticities across suppliers. These supply estimates are highly correlated with our quantity discounts. One note, that estimation explicitly assumed monopolistic competition and CES demand. Together, these factors can explain about half the variability in

Table 5: Quantity Discounts: Relationship Level Heterogeneity

	(1) log p		(7) log p
Log(Q) x Related Party	-0.00571	Log(Q) x Buyer-HS-Country Share	-0.000000899
	(0.000171)		(0.000000154)
Within r^2	0.000289	Within r^2	0.00000875
	(2)		(8)
Log(Q) x Input Share	-0.0147	Log(Q) x log(Q Seller)	0.0172
	(0.000222)		(0.0000318)
Within r^2	0.00114	Within r^2	0.00749
	(3)		(9)
Log(Q) x Output Share	0.0237	Log(Q) x log(Q Buyer)	0.0166
	(0.000288)		(0.0000316)
Within r^2	0.00175	Within r^2	0.00708
	(4)		(10)
Log(Q) x log(Q Relationship)	0.0171	Log(Q) x HHI Seller	-0.00384
	(0.0000306)		(0.000347)
Within r^2	0.00803	Within r^2	0.000317
	(5)		(11)
Log(Q) x log(Relationship Length)	0.00401	Log(Q) x HHI Buyer	0.00326
	(0.0000521)		(0.0000319)
Within r^2	0.000155	Within r^2	0.000027
	(6)		
Log(Q) x Seller-HS-Country Share	0.0000581		
	(0.00000261)		
Within r^2	0.00000128		
Fixed Effects		Fixed Effects	
Buyer-Seller-HTS10	x	Buyer-Seller-HTS10	x
HTS6-log q Slopes	x	HTS6-log q Slopes	x

Notes: See text for details. This displays heterogeneity in quanity discounts as a function of various attributes of a relationship, buyer, or seller.

quantity discounts.

In particular, products that are highly substitutable or homogeneous (a high σ/ a low σ^{-1} exhibit smaller quantity discounts. Does production require easier scaling, or is this reflective of low product costs?

Quantity discounts appear to reflect underlying supply and product characteristics, not exercises of differential market power, either between sellers and buyers. This view

largely holds to broad findings from IO.

That largely leaves models with fixed costs for searching/switching costs, and transactions or models with declining variable costs. This lines up well with qualitative evidence from operations research surveys (Mason, 1939; Schotanus, Telgen, and de Boer, 2009; Munson and Rosenblatt, 1998).

The literature and even questionnaires of firms finds two primary categories for quantity discounts. One is a pure reflection of costs, in particular scale economies (perhaps derived from fixed transaction costs) on the seller's side. However, there could also be a reflection on search costs on the buyer-side or some other mechanism that could lead to second degree price discrimination and distortions with macroeconomic effects (Bornstein and Peter, 2024). There is market power, but all sellers exert the same degree of market power in an industry. The question is what is the relative effect of search costs here (Marshall, 2020).

Seller Returns to Scale We first consider returns to scale on the seller side. The seller's cost story is consistent with identical quantity discounts for shipments within and outside a firm (as seen in Table 5, regression 1). If prices for within-firm transactions are representative of cost[31], then economies of scale are perfectly representative of returns to scale and reflect cost decreases with a transaction's scale.

One characteristic of trade data is that shipping and freight costs are considered independently of a transaction cost. With small exception, freight costs reflect all the buyers and sellers in a market, with most buyers and sellers acting as price takers. Are these quantity discounts also available in shipping? While there may be market power in shipping. A reasonable expectation is that: $\log(p_t^s) = \gamma^s \log(q_t) + FE_r + \epsilon_t$, where the price of shipping p_t^s for transaction t within relationship r is computed at the shipping cost divided by the number of unit shipped q_t. [32]

In Table 6, regression (1), we show that freight costs exhibit a scale elasticity of -0.369 within a relationship. That a doubling of an order's size [33], leads to a 37% decrease in unit shipping costs, between orders between the same buyer and seller within a product category.

[31] As is consistent with one of the two methods of determining valuations of transactions

[32] One note here is that many shipping costs computed at the HS level by the US Census are imputed, as goods across HS-codes can be bundled together in a single shipment or container. This imputation process, while relatively transparent, still does add some measurement error to the results. Limited subset to single-good shipments shows similar relationships.

[33] We leave the study of shipping scale to future work, tying in with both market power and economies of scale, as in Ignatenko (2020); Ganapati, Wong, and Ziv (2024).

Table 6: Shipping Cost Scale

	(1)
	$\log(p_{Shipping})$
log(Q)	-0.427
	(0.00854)
R^2	0.937
R^2 (within)	0.204
Fixed Effects	Relationship

Notes: See text for details.

Second-degree Price Discrimination While our evidence is predominately in support of transaction-level scale economies, second degree price discrimination may be another source of quantity discounts. Broadly speaking, the purpose of second degree price discrimination is to segment purchases to take advantage of varying downstream demand elasticities, whereby larger purchasers have more sensitive demand.

Mechanically, results rely on some level of (a) imperfect competition and (b) varying elasticities of demand. However, as shown in Table 5 and 7, typical measures of market power (in the cross-section) are not particularly strong. However, markets with large quantity discounts have products with weaker substitutability. How should we consider this apparent contradiction?

One extremely appealing solution comes from the transaction and search cost literatures.

While fixed costs in selling can generate economies of scale, as do fixed costs in searching for sources, when combined with hard to substitute products - regardless of standard measures of market power. Notably, it appears possible that a fully structural model can fit this data either way absent data on either buyer search costs or supply side costs, as shown by XXXX.

Broadly speaking, the quantity discount schedule doesn't vary across buyers and sellers within product codes. Indicating either a model of increase returns to scale on the seller-side, or fixed search costs on the buyer side. [34]

Having established the ubiquity and prevalence of quantity discounts in the short and medium run and the fact they are driven by attributes of the underlying product and

[34]think of a model of heterogeneous buyers, each receiving a draw to buy q units and facing fixed costs f to discover a price and verify quality. with a large mass of sellers, they will post a two-part tariff, or equivalent to capture sellers with low q draws, who will be unwilling to pay an additional f draw. See Marshall (2020) for an example.

Table 7: Quantity Discounts: Market Level

	(1)	(2)	(3)	(4)	(5)
	Quantity Discount γ_v				
HHI Seller	-0.0522				
	(0.0796)				
HHI Buyer		0.0027			
		(0.0679)			
log(Buyers)			0.0032		
			(0.00484)		
log(Sellers)				0.00648	
				(0.00439)	
log(Pairs)					0.00856
					(0.00435)
R-squared	0.001	0	0.001	0.003	0.005

	(6)	(7)	(8)	(9)	(10)
	Quantity Discount γ_v				
log(Transactions)	0.0163				
	(0.00357)				
log(Value)		0.0184			
		(0.00344)			
Omega			0.226		0.369
			(0.0133)		(0.0196)
Sigma^{-1}				-0.04	-0.366
				(0.0236)	(0.0268)
R-squared	0.027	0.037	0.225	0.003	0.395

Notes: See text for details.

the nonlinear fixed costs in either production or acquisition, the question is, "does this matter". We answer this in three ways.

5 Aggregate Implications

We now consider the implications for economics of scale at the transaction level. First considering aggregate supply elasticities and estimates of economies of scale (such as those in Lashkaripour and Lugovskyy (2023), Bartelme, Costinot, Donaldson, and Rodriguez-Clare (2019), and Farrokhi and Soderbery (2020)). Second, we consider how to interpret pass-through estimates, in our case looking at US tariff pass-through. Lastly, we consider the productivity and production function literature and highlight the importance of correcting for input prices.

5.1 Aggregation

Scale economies can take many forms and have significant aggregate implications on welfare computations and optimal policy. Lashkaripour and Lugovskyy (2023) find scale economies at the firm level and consider their implications for optimal trade policies. Bartelme et al. (2019) have scale economies at the sectoral level and find implications for these policies on optimal industrial policies. While most estimates of production functions estimation do not find scale (De Ridder, Grassi, and Morzenti, 2024), they often do so under the implication that realized input (or output) prices are not a function of the quantity bought (a major exception being Morlacco (2019), who relaxes this for just imported inputs at an aggregate firm-wide level).

A large question is what are the roots of scale economies? (See the urban economics literature, as in Duranton and Puga (2004) , often finding scale economies, but largely agonistic on the root causes).[35]

However to make progress, we attempt to see how much of aggregate variation in price between firm is due to transaction-level scale, as opposed to variation at the firm level. That is, are Amazon's input prices low since they buy a lot overall? Or because each shipment they buy is large and there are transaction-level quantity discounts. Do these elasticities or discounts reflect quantity discounts at a higher level?

Fundamentally, our idea is to separate the bundle of price and quantity to a standardized price for a single quantity. The standardized price, nets out a transaction-level quan-

[35]Scale economies are also compatibility with reduced for evidence that big firms pay less as in Kugler and Verhoogen (2012), as they have different qualities, and they often buy more inputs, with the marginal inputs at very low quantities.

tity discount. We can then consider, net of this, are there quantity or scale effects at higher levels of aggregation, or are they mostly reflective of our transaction-level effects?[36]

From the fixed effects in the regression, we can take for every transaction the following object, [37]

$$\tilde{p}_{i\to j,t,v} = \exp\left(\gamma_{i\to j,v} + \xi_{i\to j,t,v}\right), \tag{12}$$

These fixed effects are essentially a buyer-seller-product quality adjusted price, after netting out transaction size, which are essentially just the residuals after netting out transaction level scale. We then collapse the data to the i-j-v level:

$$\log \tilde{p}_{i\to j,v} = \log \sum_t q_{i\to j,t,v} \tilde{p}_{i\to j,t,v} \tag{13}$$

After considering how much price variation is accounted for by at the lowest level of aggregation, how much does the frequency and aggregate size of buying affect pricing?

We can then run regressions on this:

$$\log \tilde{p}_{i\to j,v} = \Theta \log q_{i\to j,v} + FE_{i,v} + FE_{j,v} + \epsilon_{ij} \tag{14}$$

In addition we can compute:

$$\bar{p}_{i\to j,t,v} = \exp\left(\xi_{i\to j,t,v}\right), \tag{15}$$

This is simply the start of a buyer-agnostic price schedule, enabling us to recover a quantity-agnostic price:

$$\tilde{p}_{i\to j,v} = \sum_t \bar{p}_{i\to j,t,v} * q_{i\to j,t,v} / \sum_t q_{i\to j,t,v}. \tag{16}$$

We can then run regressions to see if on this, along the vein of the decomposition in 2 to understand what correlates to aggregate relationship-level prices, as those studied by Alviarez et al. (2023) and Kamal and Sundaram (2016):

$$\log \tilde{p}_{i\to j,v} = \theta \log q_{i\to j,v} + FE_{i,v} + FE_{j,v} + \epsilon_{ij}. \tag{17}$$

Table 8 highlights these regressions in two forms. The first using raw price data and the second using our residual price $\tilde{p}$, which controls for the transaction-level quantity

[36] For example, feeding back, it may be that an increase in buyers and seller, feedback to reducing fixed costs, leading to changes in both prices and quantity discounts

[37] In a forthcoming appendix we micro-founded elements of demand using an inventory method, where Demand at the transaction level is stochastic. (intensive margin) governed purely by physical constraints.

Table 8: Relationship-Level Cross Sectional Price Differences

	(1) $\log p$	(2) $\log \tilde{p}$	(3) $\log p$	(4) $\log \tilde{p}$	(5) $\log p$	(6) $\log \tilde{p}$
$\log q$	-0.221	-0.074	-0.151	-0.025	-0.136	-0.020
	(0.002)	(0.001)	(0.002)	(0.001)	(0.003)	(0.001)
r^2	0.391	0.236	0.766	0.715	0.906	0.89
Within r^2	0.223	0.052	0.154	0.009	0.135	0.006
Fixed Effects	Country-HTS10		Seller-HTS10		Buyers-HTS10, Seller-HTS10	

Notes: See text for details.

discount. Column (1) shows that with product and source fixed effects, the aggregate quantity between a seller and buyer, accounts for 23% of the price variation, with an aggregate elasticity θ = -0.22. Column (2) shows that accounting for transaction level quantities implies that aggregate quantities only explain 5% of the price variation. However this accounting implicitly makes assumptions about the unique nature of a product. Columns (3) and (4) allow for seller fixed effect[38]. Results are broadly similar, however quantities now only explain 1% of the residualized price.

Columns (5) and (6) include the most demanding set of fixed effects, buyers and seller at the HTS 10-digit product code. Column (5) shows that there is a relatively significant correlation between raw prices and aggregate quantities, with θ = -.137, explaining 13.5% f the variation. However, when controlling for transaction level quantities in (6), there is an extremely small effect, with almost no relationship between the quantity bought and the price paid.

In unreported regressions, we additionally run regressions on importer and exporter shares and volumes, as in section 2, but find they have minimal additional explanatory power as before, once relationship-level quantities are controlled for.

However, these regressions are purely a decomposition, they aren't reflecting a supply curve, but rather the equilibrium output of millions of relationships in the cross-section. To do so, we turn to a first difference strategy, that exploits changes in underlying demand to estimate a relationship level supply curve, both with and without accounting for economies of scale in individual transactions.

[38]The forthcoming appendix includes buyer fixed effect

Table 9: Bilateral Price Change - International Trade

	(1) $\Delta \log p$	(2) $\Delta \log \tilde{p}$	(3) $\Delta \log p$	(4) $\Delta \log \tilde{p}$	(5) $\Delta \log p$	(6) $\Delta \log \tilde{p}$	(7) $\Delta \log p$	(8) $\Delta \log \tilde{p}$
$\Delta \log q$	-0.209	-0.0225	-0.222	-0.0229	-0.193	-0.0184	-0.115	-0.00114
	(0.005)	(0.001)	(0.006)	(0.001)	(0.006)	(0.001)	(0.04)	(0.002)
N	230000	230000	230000	230000	148000	148000	148000	148000
r^2	0.353	0.187	0.527	0.419	0.527	0.431	0.153	0.0005
Within r^2	0.197	0.006	0.21	0.006	0.182	0.004		
Fixed Effects	Country-HTS10				Seller-HTS10			
First Stage F							1000	
Instruments							X	X
IV Sample					X	X	X	X

Notes: This comes from a regression of change in aggregates quantities on a measure of price. Odd columns use aggregate prices. Even columns follow the correction in equation X, and adjust each transaction's price for the quantity sold, using the aggregate relationship from XXX. Columns (1) through (6) use fixed effects in increasing order of stringency. The last two columns instrument for the change in quantity using the instrument detailed in the text.

5.1.1 First Differences

How do we identify a relationship-level supply curve? While extremely difficult to do in the cross section, we will adapt the strategy from 3.2 to identify a simplified inverse supply elasticity θ using a combination of first differences over a medium time span (one year from 2011 to 2012), as well as an shift-share style instrument to correct for any residual endogeneity.

We first control, as earlier for buyer-seller-product fixed effects using a first difference operator. Secondly, we control for a seller-product fixed effect to account for any supply side shocks, our major threat to identification. In the forthcoming appendix, we run a variation where we also do control for buyer-product fixed effects, therefore not relying on a buying firm shock, but rather a shock to the specific bilateral relationship. Results are quantitatively similar.

We can then run regressions on this:

$$\Delta \log \tilde{p}_{i \to j,v} = \theta \Delta \log q_{i \to j,v} + FE_{i,v} + \epsilon_{ij}, \tag{18}$$

where all $\tilde{p}$ is computed after netting out the effect of transaction level quantities as before. As a comparison, we also continue to run regression using the raw price data, $\log p$, as a comparison to see if there are economies (or dis-economies of scale) at the relationship

level, how much is accounted for by the underlying transaction level quantity discounts?

5.1.2 Instrumentation

As before, we must distinguish from supply side shocks, as we want to use demand variation to identify the supply line (cite Berry et al). While, supply-product trends control for supply shocks at the seller level, there may be bilateral shocks from bilateral investment outcomes. We use aggregate demand shocks that differentially affect different buyers in downstream industries (in the forthcoming appendix we do robustness to further consider potential second order effects). One set of demand shocks is how much more did a buyer change of a particular product from other suppliers, weighted by how much they bought from other suppliers in the initial period.

In aggregate we find little effect. Further we can tie into the bilateral monopoly research too and use the further network-structure of demand to instrument, but results are largely similar.

5.1.3 Domestic Trade

There do not appear to be aggregate scale economies (or cross-sectional or time series effects of changes in aggregate order volumes and indication of unilateral or bilateral market power) in international imports. What about domestic trade? We conduct the same exercise, first a decomposition, then a regression controlling for supply side effects with an instrument.

There are two downsides to the domestic trade data. First, it is not a time series, severely limiting both controls, as well as potential instruments. However, even in the cross-section, we can control for relationship fixed effects. Furthermore, the domestic data has one advantage over the trade data, for a single seller, we see a wider swath of buyers and can directly control for some bilateral effects.

In Table 10. column (1) and (2), we show that while a substantial portion of the variation in aggregate prices $\log p$ can be correlated with total volumes $\log q$, once transaction level discounts are stripped out, there are no correlations. Columns (3) and (4) add significant buyer-product and seller-product fixed effects and show that this relationship is unchanged.

However to show causality, we need a shifter of demand, not some potential bilateral shifter of supply. Columns (5) and (6) implement a simplified version of the instrument used in the import data, at the destination, how much overall demand exists, without including the focal source? Essentially a location that buys lots of car parts will have

more demand for car parts, even excluding the focal source location. Further to control for geographic proximity, we control for distance (as measured using road distance as computed by the US Census and Department of Transportation).

Without controlling for transaction level quantity discounts, $\theta = -.11$, however controlling for transaction level quantity discounts, $\theta = .02$, close to constant returns to scale.

Table 10: Decomposing Relationship Prices - Domestic Trade

	(1) $\log p$	(2) $\log \tilde{p}$	(3) $\log p$	(4) $\log \tilde{p}$	(5) $\log p$	(6) $\log \tilde{p}$
$\log q$	-0.15	0.0188	-0.15	0.0288	-0.112	0.024
	(0.00607)	(0.00634)	(0.00609)	(0.00628)	(0.00605)	(0.00609)
log(Distance)			-0.0329	-0.0161	-0.0154	0.0139
			(0.00310)	(0.00280)	(0.00353)	(0.00384)
r^2	0.823	0.947	0.86	0.957	0.112	0.12
Within r^2	0.123	0.00219	0.12	0.00481		
First Stage F-Stat					396.6	396.6
Fixed Effects	Seller, SCTG5		Buyer-SCTG5, Seller-SCTG5			
Instruments					X	X

Notes: This comes from a regression of change in aggregates quantities on a measure of price. Odd columns use aggregate prices. Even columns follow the correction in equation X, and adjust each transaction's price for the quantity sold, using the aggregate relationship from XXX. Columns (1) through (6) use fixed effects in increasing order of stringency. The last two columns instrument for the change in quantity using the shift-share instrument detailed in the text. Standard errors are clustered by SCTG product.

5.1.4 Aggregate Effects

After accounting for transaction-level scale economies, do we find aggregate relation-level scale economies in trade? (Aligning with Lashkaripour and Lugovskyy (2023)) or those even at higher levels of aggregation (aligning with Bartelme et al. (2019)). Furthermore can we explain aggregate material scale, as in Atalay (2014) or Davis et al. (2013)?

The answer is nuanced. There still appear to be scope for aggregate scale economies, but accounting for transaction size greatly reduces their magnitudes, However on the flip-side, as a first approximation, aggregate scale economies mechanically absorb transaction-level scale economies - allowing to a first approximation - macroeconomic analysis to be done. Essentially, a very large portion of the aggregate scales economies measured are simply a reflection of transaction-level scale economies.

Scale seems to matter for aggregation, one of the clearest places aggregation occurs is in the pass-through of tariffs. We now turn to a case study of the imposition of tariffs and taxes on US importers by the US government in 2018.

5.2 Tariff Pass Through

One important aggregate object related to aggregation is the rate of pass through (which is some models is one of a a handful sufficient statistics necessary to compute welfare and incidence for policy choices). Our primary extension of this literature is accounting for the size of transaction changes the interpretation of aggregate pass-through statistics, as well as allows for more accurate pass-through statistics to be computed. [39]

Our approach is to treat a combination of product and quantity as a single bundle. Specifically, we consider a product not as as combination a source-destination-variety-quantity bundle, instead of the simpler source-variety or even a source-destination-variety. Essentially, we empirically recover issues with reduced form versus more structural estimates of pass-through (MacKay, Miller, Remer, and Sheu, 2014).

We view data from 2017-2019, covering the bulk of the US-China Tariff war period, before the effects of COVID-19 complicate results. (cite here). We do find that at both the aggregate and transaction level, we can match the initial findings of the literature (Fajgelbaum, Goldberg, Kennedy, and Khandelwal, 2020). That is considering a variety, US tariffs on imports exhibit close to full pass-through. This not only is at the aggregate-level, but also at the bilateral buyer-seller-variety level (Handley, Kamal, and Monarch, 2025). However there are large quantity responses, not just from the aggregate demand curve, but also in the size of individual transactions.

At the transaction level, we start off with the relationship:

$$\log p = \log \tilde{p} + \gamma \log q., \tag{19}$$

where the price is inclusive of tariffs.[40]

[39] Unlike our main regressions, we now alter our definition, while earlier, prices were all demeaned, here we keep them at their initial levels. Additional, for the aggregate regressions, we do not drop 'singletons', see text below for full details.

[40] An alternative, but equivalent computation is writing

$$\begin{aligned} \log p_t &= \log \bar{p} + \log(1+\tau) + \gamma \log q_t \\ \frac{d \log p_t}{d \log(1+\tau)} &= \frac{d \log \bar{p}}{d \log(1+\tau)} + 1 + \gamma \frac{d \log q_t}{d \log(1+\tau)} \end{aligned}$$

Now if we assume that $\bar{p}$ is invariant to tariffs, then the first term on the right hand side drops out.

For any given transaction t from seller i to buyer j at date d, as shown in section 3.2, there is a scale pricing elasticity γ. When considering tariffs, tariffs have two log-separable effects on prices:

$$\frac{\partial \log p}{\partial \log (1+\tau)} = \frac{\log \tilde{p}}{\partial \log (1+\tau)} + \gamma \frac{\log q}{\partial \log (1+\tau)}. \tag{20}$$

The first element on the right hand side is the effect of a tariffs on a scale-independent price. The second element is the effect of a tariff on the scale of an individual transaction.

5.2.1 Transaction-Level Passthrough

Looking at data over the 2017-2019 period on all tariffs imposed by the US on its own imports, Table 11 show the results of considering equation 20. We consider prices, not only within a buyer-seller-product relationship, but relative to product-time trends (to partially account for the supply shock effect in Gertler and the diversions seen in Goldbery 2024 AER Insights)

Column (1) highlights nearly full pass-through of prices looking at continuing relationships. A 1% increase in tariffs translates to a .96% increase in prices. Column (3) considers the role on order size. A 1% increase in tariffs, leads to a 0.6% decrease in the average order size. Column (4) highlights our mechanism, where if order size stayed constant, supply elasticities would indicate that passthrough was just 77.5%, far below unity. The last column (5), shows simply that the OLS adjustment of simply including order size in the regression acts as a rough approximation for the two stage correction.

In the forthcoming appendix we include the intermediate step, aggregating transactions to a pre and post tariff level (excluding the month of implementation), and show how these results translate to that level of aggregation. Additionally, we can show that γ itself is invariant to tariff rates.

5.2.2 Short-Run Aggregate Pass-through

Moving on to aggregation adds a handful of twists. First are the number of orders. Second are the changes in the structure of relationships. Focusing on aggregation at the origin-product level, the same level of aggregation most public use data has, we outline two decompositions. The first decomposition leverages the empirical supply elasticities from Section 3 to compute a counterfactual pass through if quantities remained constant, but requires access to transaction-level data to compute. The second mechanically preforms a first order approximation, using aggregate data and finds a quantitative and qualitatively

Table 11: Transaction-level Tariff Pass-Through: Within Relationships

	(1) $\log p$	(2) $\log Value$	(3) $\log q$	(4) $\log \tilde{p}$	(5) $\log p$
log(1+Tariffs)	0.956	0.357	-0.600	0.775	0.775
	(0.00460)	(0.00672)	(0.00778)	(0.00396)	(0.00396)
$\log q$					-0.302
					(0.000669)
r^2	0.965	0.725	0.912	0.000662	0.974
Within r^2	0.000746	4.86E-05	0.000103	0.000662	0.261
Fixed Effects	Relationship, HTS10-Time				

similar results.

For the first method, we initially adjust prices, netting our quantity discount out, so at the transaction level we recover a standardized price for a unit[41],

$$\tilde{p}_{i\to j,t,v} = \frac{p_{i\to j,t,v}}{q^{\gamma}_{i\to j,t,v}}, \tag{21}$$

where γ is recovered through the mechanism in section 3.2.[42]

With this, we aggregate across all buyers and sellers, where shipment of product code v originates in country o at date d, aggregated at the monthly level:

$$\tilde{p}_{o,d,v} = \sum_{i\in o,t\in d} \tilde{p}_{i\to j,t,v} q_{i\to j,t,v} / \sum_{i\in o,t\in d} q_{i\to j,t,v}. \tag{22}$$

With this we decompose the aggregate price for a product as:

$$\log p_{o,d,v} = \log \tilde{p}_{o,d,v} + \log \tilde{q}_{o,d,v}, \tag{23}$$

where we call $\tilde{q}$ the aggregate quantity shift. Now due to Jensen's inequality, the logarithm of a sum is not the sum of a logarithm as log is a convex function, thus requiring the use of detailed transaction level data in construction.

Table 12 runs regresses each of these components, both the left and right hand sides on changes in applied tariffs. Column (1) replicates the baseline findings, pass through,

[41] Going back to equation XXX, we opt to leave the deviation in of the price, however results are robust to this step

[42] We alternatively use γ_v at the variety level to the same effect

not just at the transaction level, but the product-origin level is effectively unity. Columns (2) and (3) replicate the decomposition. Column (2) indicates that passthrough of a

As an alternative, we can bound our results using a much more straightforward decomposition,

$$\log p_{o,d,v} = \log pq_{o,d,v} - \log q/T_{o,d,v} + \log T_{o,d,v}, \tag{24}$$

where T denotes the number of transactions. In particular the second term is informative of the level of pass-through netting out quantity changes, armed with our γ from Section 3.2.

Table 12 in columns (4)-(7) regresses tariff rates on each of these components. Column (4) indicates a 1% increase in tariffs leads to a 1.5% in payments. Column (5) shows that this is driven by a 2.5% decrease in quantity. Crucially, column (6) shows that a 1% increase in tariffs leads to a -1.8% decrease in average order size. As an approximation,

$$\frac{d\log \tilde{p}_{o,d,v}}{d\log(1+\tau_{o,d,v})} \approx -\gamma \frac{d\log (q/T)_{o,d,v}}{d\log(1+\tau_{o,d,v})} = -1.881 * .30 = -0.54.$$

Essentially, recovering column (2) to a first order approximation. For a consistent order size, tariffs are only passed through at the aggregate level at 50%.[43]

So why is this different that the 78% pass-through from Table 11? Column (7) gives some clues, the number of orders also falls. We additionally do not account for what orders are from continuing firms, which are from entering firms, and which are from exiting firms. For that we turn to another decomposition.

Table 12: Aggregate Pass-Through, Full Period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\log p$	$\log \tilde{p}$	$\log \tilde{q}$	$\log pq$	$\log q$	$\log(q/T)$	$\log T$	$\log p$
log(1+Tariffs)	1.01	0.583	0.427	-1.493	-2.502	-1.881	-0.621	-0.18
	(0.0376)	(0.0149)	(0.0280)	(0.0395)	(0.0548)	(0.0472)	(0.0229)	(0.0271)
$\log q$								-0.475
								(0.0003)
r^2	0.908	0.878	0.902	0.87	0.9	0.892	0.904	0.952
r^2 within	0.0003	0.0006	0.0001	0.0006	0.0009	0.0007	0.0003	0.481
Fixed Effects	HTS10-Time, HTS10-Country Origin							

[43]Column (8) shows that unlike with transaction level data, you must know γ directly and cannot estimate in this step without further information

5.2.3 Medium-Run Aggregate Pass-through

In aggregate, pass through is the combination of two forces, the pass-through of continuing firms, and the net effect of entrants and exiters.

Here we focus on the longer run difference. With the bulk of the 2018 tariffs implemented in September 2018 (and slowly adjusted until January 2019), we consider the medium run difference between June 2018 and June 2019.

Working backwards, we run effectively the same decomposition as in the previous subsection, but in first differences, as there are only two periods. Panel 3, runs regressions at the aggregate level, finding that tariff pass-through is approximately 1.09 at this horizon, with 0.57 attributed to price effects and 0.52 to a downward shift in quantity.

Panel 2, only does this with the net effect of entrant and exiting trading pairs. Aggregate pass-through, under the assumption of comparability implies that pass-through is more than full (1.2) for this set of firms, with only a small quantity effect (.25), with the price effect being full.

Panel 1, does this for continuing trading partners. In this case, aggregate pass-through is only .78. The price effect is effectively 0 (.06), with a large quantity effect (.72), as volumes shift downward. (The sign is opposite the actual change in quantities).

In all three panels, this effect can be appropriated as above, using the exact decomposition and our estimate of γ.[44]

Overall, using a more conventional price index, as opposed to a price index that is not based on quantities-adjusted prices, would show substantially less than full pass-through of tariffs, with the net effect on the order of .50, with the difference driven by substantial decreases in the average order size, as well as a decrease in the number of orders.

What remains is to rationalize this pass-through in terms of market conduct and the presence of both the measured supply and demand elasticities.

[44] An alternative exact decomposition is:

$$\begin{aligned}
\log\left(\frac{p'}{p}\right) &= \log\left(\frac{w_c' p_c' + w_{entrant} p_{entratns}}{w_c p_c + w_{exit} p_{exit}}\right) \\
&= \log\left(\frac{w_c' p_c'}{w_c p_c}\right) + \log\left(1 + \frac{w_{entrant} p_{entratns}}{w_c' p_c'}\right) - \log\left(1 + \frac{w_{exit} p_{exit}}{w_c p_c}\right) \\
&= \log\left(continuers\right) + \log\left(entrants\ rel\ continue\right) - \log\left(exit\ rel\ continues\right)
\end{aligned}$$

Table 13: Decomposing Medium Run Tariff Pass Through

Panel 1: Continuing Partners

	(1) $\Delta\log p$	(2) $\Delta\log Value$	(3) $\Delta\log q$	(4) $\Delta\log(q/T)$	(5) $\Delta\log T$	(6) $\Delta\log\tilde{q}$	(7) $\Delta\log\tilde{p}$
Δlog(Tariffs)	0.781	0.522	-0.259	-0.253	-0.00628	0.721	0.0595
	(0.196)	(0.259)	(0.318)	(0.263)	(0.182)	(0.151)	(0.0839)
r^2	0.153	0.159	0.156	0.154	0.159	0.157	0.147
Within r^2	0.000324	8.24E-05	1.36E-05	1.88E-05	2.43E-08	0.000464	1.02E-05

Panel 2: Exiting and Entering Partners

	(1) $\Delta\log p$	(2) $\Delta\log Value$	(3) $\Delta\log q$	(4) $\Delta\log(q/T)$	(5) $\Delta\log T$	(6) $\Delta\log\tilde{q}$	(7) $\Delta\log\tilde{p}$
Δlog(Tariffs)	1.208	-2.494	-3.703	-3.209	-0.494	0.254	0.954
	(0.305)	(0.323)	(0.453)	(0.382)	(0.190)	(0.225)	(0.122)
r^2	0.128	0.154	0.145	0.138	0.163	0.13	0.135
Within r^2	0.000289	0.0011	0.00123	0.0013	0.000124	2.36E-05	0.00113

Panel 3: Aggregate

	(1) $\Delta\log p$	(2) $\Delta\log Value$	(3) $\Delta\log q$	(4) $\Delta\log(q/T)$	(5) $\Delta\log T$	(6) $\Delta\log\tilde{q}$	(7) $\Delta\log\tilde{p}$
Δlog(Tariffs)	1.087	-0.748	-1.835	-1.902	0.0666	0.522	0.565
	(0.229)	(0.247)	(0.339)	(0.288)	(0.152)	(0.169)	(0.0930)
r^2	0.113	0.137	0.128	0.12	0.145	0.117	0.117
Within r^2	0.000304	0.000124	0.000395	0.000588	2.61E-06	0.000128	0.000499

Fixed Effects:	HTS10 - Year - Month

Notes: TBD

5.2.4 Reconciling Theory and Transaction-level Passthrough a la Weyl and Fabinger (2013)

Consider industry-wide cost pass-through with symmetric firms. Weyl and Fabinger (2013) derive the following general expression in this case:

$$\frac{dp}{dt} = \frac{1}{1 + \frac{\epsilon_d - \theta}{\epsilon_s} + \frac{\theta}{\epsilon_\theta} + \frac{\theta}{\epsilon_{ms}}} \tag{25}$$

where θ is a conduct parameter (ranges from 1 for monopoly to 0 for perfect competition), ϵ_d is the elasticity of demand, ϵ_s is the elasticity of supply, ϵ_θ is the elasticity of the conduct parameter, and ϵ_{ms} is a measure of demand curvature ($\frac{1}{\epsilon_{ms}} = 1 + \frac{qp''(q)}{p'(q)} =$

$1 + \frac{(\frac{pq''(p)}{q'(p)})}{\epsilon_d}$).

With symmetric, homogeneous Cournot, this becomes

$$\frac{dp}{dt} = \frac{1}{1 + \frac{\epsilon_d - HHI}{\epsilon_s} + \frac{HHI}{\epsilon_{ms}}} \tag{26}$$

where HHI is the Herfindahl index ($HHI = 1/N$). Note that $\epsilon_d > 0$, but ϵ_s and ϵ_{ms} are not sign restricted.

In the case of $\epsilon_s < 0$, to get $\frac{dp}{dt} > 0$ requires that $\epsilon_{ms} > 0$ and that $\frac{1}{\epsilon_{ms}}$ is sufficiently large.

Note that $\frac{1}{\epsilon_{ms}}$ = 1 - curvature term of Mrázová and Neary (2017). As an aside, in monopolistic competition, the firm's second-order condition requires the curvature term < 2. Thus, monopolistic competition implies that $\frac{1}{\epsilon_{ms}} > -1$. In symmetric oligopoly, the second-order condition requirement loosens to the curvature term $< 2N$, implying $\frac{1}{\epsilon_{ms}} > 1 - 2N$.

So, in general these conditions are a priori possible. How would one rationalize this in a quantitative model? The tricky part is getting a demand system paramaterized to satisfy these conditions, as standard demand systems are not flexible with respect to the curvature of demand. Here I show that an extension to the homothetic translog demand system Christensen, Jorgenson, and Lau (1975) is flexible with respect to both the first order and second order properties of demand (locally).

Omitting the cross-price terms for ease of exposition, the own price elasticity of demand $\varepsilon^{q_j}_{p_j}$ is given by:

$\varepsilon^{q_j}_{p_j} \equiv -\frac{\partial q_j}{\partial p_j}\frac{p_j}{q_j} = \frac{-b_{jj} - 2b_{jjj} log p_{jt}}{s_{jt}} + 1$

The Mrázová and Neary (2017) measure of curvature (denoted ρ_j) is the ratio of the own price convexity of demand to the own price elasticity of demand and is given by:

$$\rho_j \equiv -\frac{\frac{\partial^2 q_j}{\partial^2 p_j}\frac{p_j}{\partial q_j / \partial p_j}}{\varepsilon^{q_j}_{p_j}} = \frac{-3\varepsilon^q_p - \frac{2b_{jjj}}{s_{jt}} - 1}{(\varepsilon^q_p)^2}$$

In this case, $\frac{1}{\epsilon_{ms}} = 1 + \frac{3\varepsilon^q_p + \frac{2b_{jjj}}{s_{jt}} + 1}{(\varepsilon^q_p)^2}$

Thus, b_{jjj} needs to be positive and sufficiently large.

[[Forthcoming]]

5.3 Estimating Production Functions, Markups and Markdowns

[[Forthcoming]]

6 Other Applications

6.1 Computing Price Indices

International price indices are often constructed using realized prices, rather than prices that control for units. (The BLS CPI and PPI are constructed using surveyed prices for a specified unit, but the trade price indices are derived from aggregate trade data). Conditional on a scale elasticity, it is straightforward to makes these price indices consistent, as in our pass-through estimates.

6.2 Shock Propagation, Bullwhip Effects, and Business Cycles

The propagation of shocks in a supply chain can be amplified due to returns to scale in both procurement and sales. An positive demand shock, may be felt upstream and firms successively increase their order sizes, with falling prices. A negative demand shock may work the other way around.

7 Conclusion

This paper shows the empirical relevance of quantity discounts at the transaction level and their importance for firm and relationship-level economies of scale. These discounts appear to largely reflect unobserved scale economies and not market power. Furthermore, these scale economies can parsimoniously explain the pass through of shocks, both tariffs and demand shocks.

Going beyond this paper (see Ganapati 2023b), the appropriate use of quantity discounts matters for any situation that uses material usage as an undistorted input, for example to measure distortions in output and labor markers (Yeh, Macaluso, and Hershbein, 2022). In a macroeconomic context, such quantity discounts matter for papers that use aggregate input-output tables, or relies on a game where upstream seller prices are taken as a given by downstream firms (Baqaee and Farhi, 2020). Furthermore, harking to (Buchanan, 1952), "If the buyer exerts any control over price, no demand curve can be derived since quantity demanded then depends on the whole market offer, not on price alone". While this may not matter for studies of non-negotiable final consumer goods, nearly all other stages of the economy seem to exhibit a degree of upstream price negotiation.

References

ALESSANDRIA, G., J. P. KABOSKI, AND V. MIDRIGAN (2011): "US trade and inventory dynamics," *American Economic Review*, 101, 303–307.

ALVIAREZ, V. I., M. FIORETTI, K. KIKKAWA, AND M. MORLACCO (2023): "Two-sided market power in firm-to-firm trade," Tech. rep., National Bureau of Economic Research.

AMITI, M., S. J. REDDING, AND D. E. WEINSTEIN (2019): "The impact of the 2018 tariffs on prices and welfare," *Journal of Economic Perspectives*, 33, 187–210.

ANTRAS, P. AND E. HELPMAN (2004): "Global sourcing," *Journal of political Economy*, 112, 552–580.

ARROW, K. J., T. HARRIS, AND J. MARSCHAK (1951): "Optimal inventory policy," *Econometrica: Journal of the Econometric Society*, 250–272.

ATALAY, E. (2014): "Materials prices and productivity," *Journal of the European Economic Association*, 12, 575–611.

BAQAEE, D. R. AND E. FARHI (2020): "Productivity and misallocation in general equilibrium," *The Quarterly Journal of Economics*, 135, 105–163.

BARTELME, D. G., A. COSTINOT, D. DONALDSON, AND A. RODRIGUEZ-CLARE (2019): "The textbook case for industrial policy: Theory meets data," Tech. rep., National Bureau of Economic Research.

BORNSTEIN, G. AND A. PETER (2022): "Nonlinear pricing and misallocation," Tech. rep., Mimeo.

——— (2024): "Nonlinear pricing and misallocation," Tech. rep., National Bureau of Economic Research.

BRODA, C. AND D. E. WEINSTEIN (2006): "Globalization and the Gains from Variety," *The Quarterly journal of economics*, 121, 541–585.

——— (2010): "Product creation and destruction: Evidence and price implications," *American Economic Review*, 100, 691–723.

BUCHANAN, J. M. (1952): "The theory of monopolistic quantity discounts," *The Review of Economic Studies*, 20, 199–208.

BURSTEIN, A. T., J. CRAVINO, AND M. ROJAS (2024): *Input price dispersion across buyers and misallocation*, Banco Central de Chile.

CHRISTENSEN, L. R., D. W. JORGENSON, AND L. J. LAU (1975): "Transcendental logarithmic utility functions," *The American Economic Review*, 65, 367–383.

DAVIS, S. J., C. GRIM, J. HALTIWANGER, AND M. STREITWIESER (2013): "Electricity unit value prices and purchase quantities: US manufacturing plants, 1963–2000," *Review of Economics and Statistics*, 95, 1150–1165.

DE LOECKER, J. AND C. SYVERSON (2021): "An industrial organization perspective on productivity," in *Handbook of industrial organization*, Elsevier, vol. 4, 141–223.

DE RIDDER, M., B. GRASSI, AND G. MORZENTI (2024): "The Hitchhiker's Guide to Markup Estimation: Assessing Estimates from Financial Data," .

DURANTON, G. AND D. PUGA (2004): "Micro-foundations of urban agglomeration economies," in *Handbook of regional and urban economics*, Elsevier, vol. 4, 2063–2117.

EDGEWORTH, F. Y. (1888): "Some new methods of measuring variation in general prices," *Journal of the Royal Statistical Society*, 51, 346–368.

FAJGELBAUM, P. D., P. K. GOLDBERG, P. J. KENNEDY, AND A. K. KHANDELWAL (2020): "The return to protectionism," *The Quarterly Journal of Economics*, 135, 1–55.

FARROKHI, F. AND A. SODERBERY (2020): "Trade elasticities in general equilibrium," *Manuscript, Purdue University*.

FEENSTRA, R. C. (1994): "New product varieties and the measurement of international prices," *The American Economic Review*, 157–177.

GANAPATI, S., J. S. SHAPIRO, AND R. WALKER (2020): "Energy cost pass-through in US manufacturing: Estimates and implications for carbon taxes," *American Economic Journal: Applied Economics*, 12, 303–342.

GANAPATI, S., W. F. WONG, AND O. ZIV (2024): "Entrepot: Hubs, scale, and trade costs," *American Economic Journal: Macroeconomics*, 16, 239–278.

GERTLER, S. (2023): "The Structural Drivers of Price and Quantity Adjustment: Insights from Tariff and Exchange Rate Pass-through," Tech. rep., mimeo.

HANDLEY, K., F. KAMAL, AND R. MONARCH (2025): "Rising import tariffs, falling exports: When modern supply chains meet old-style protectionism," *American Economic Journal: Applied Economics*, 17, 208–238.

HILLBERRY, R. AND D. HUMMELS (2013): "Trade elasticity parameters for a computable general equilibrium model," in *Handbook of computable general equilibrium modeling*, Elsevier, vol. 1, 1213–1269.

HOTTMAN, C. J., S. J. REDDING, AND D. E. WEINSTEIN (2016): "Quantifying the sources of firm heterogeneity," *The Quarterly Journal of Economics*, 131, 1291–1364.

HUMMELS, D. AND A. SKIBA (2004): "Shipping the good apples out? An empirical confirmation of the Alchian-Allen conjecture," *Journal of political Economy*, 112, 1384–1402.

IGNATENKO, A. (2020): "Price discrimination in international transportation: Evidence and implications," *Retrived from: https://www. annaignatenko. com/JMP_AnnaIgnatenko. pdf*.

KAMAL, F. AND A. SUNDARAM (2016): "Buyer–seller relationships in international trade: Do your neighbors matter?" *Journal of International Economics*, 102, 128–140.

KROLIKOWSKI, P. M. AND A. H. MCCALLUM (2025): "Tariffs and Goods-Market Search Frictions," Tech. rep.

KUGLER, M. AND E. VERHOOGEN (2012): "Prices, plant size, and product quality," *The Review of Economic Studies*, 79, 307–339.

LASHKARIPOUR, A. AND V. LUGOVSKYY (2023): "Profits, scale economies, and the gains from trade and industrial policy," *American Economic Review*, 113, 2759–2808.

MACKAY, A., N. H. MILLER, M. REMER, AND G. SHEU (2014): "Bias in reduced-form estimates of pass-through," *Economics Letters*, 123, 200–202.

MANNING, A. (2013): *Monopsony in motion: Imperfect competition in labor markets*, Princeton University Press.

MARSHALL, G. (2020): "Search and wholesale price discrimination," *The RAND Journal of Economics*, 51, 346–374.

MASON, E. S. (1939): "Price and production policies of large-scale enterprise," *The American economic review*, 29, 61–74.

MORLACCO, M. (2019): "Market power in input markets: Theory and evidence from french manufacturing," *Unpublished, March*, 20, 2019.

MRÁZOVÁ, M. AND J. P. NEARY (2017): "Not so demanding: Demand structure and firm behavior," *American Economic Review*, 107, 3835–3874.

MUNSON, C. L. AND M. J. ROSENBLATT (1998): "Theories and realities of quantity discounts: An exploratory study," *Production and operations management*, 7, 352–369.

OSTLUND, H. (1939): "The robinson-patman act and quantity discounts," *The Accounting Review*, 14, 402–409.

RAUCH, J. E. (1999): "Networks versus markets in international trade," *Journal of international Economics*, 48, 7–35.

SCHOTANUS, F., J. TELGEN, AND L. DE BOER (2009): "Unraveling quantity discounts," *Omega*, 37, 510–521.

SODERBERY, A. (2015): "Estimating import supply and demand elasticities: Analysis and implications," *Journal of International Economics*, 96, 1–17.

SOKOLOVA, A. AND T. SORENSEN (2021): "Monopsony in labor markets: A meta-analysis," *ILR Review*, 74, 27–55.

WEYL, E. G. AND M. FABINGER (2013): "Pass-through as an economic tool: Principles of incidence under imperfect competition," *Journal of political economy*, 121, 528–583.

YEH, C., C. MACALUSO, AND B. HERSHBEIN (2022): "Monopsony in the US labor market," *American Economic Review*, 112, 2099–2138.

A Appendix Forthcoming

[Theory Foundation]

[Additional Empirical Results]

[Aggregation]