Did Foreigners Pay America's Tariffs? Quantity Discounts, Scale Economies and Incomplete Pass-Through*

Sharat Ganapati
Georgetown University & NBER

Colin J. Hottman
Federal Reserve Board

Feb 2026

Abstract

Transaction-level quantity discounts are a pervasive feature of US trade, shaping both price variation and tariff incidence. Using administrative microdata, we show that these discounts reflect transaction-level scale economies rather than market power. Accounting for these micro-level economies resolves a key puzzle: while observed import prices rose one-for-one with 2018-2019 US tariffs, we show this was driven by the loss of scale economies as transaction sizes collapsed. Controlling for this scale effect, the strategic pass-through of tariffs to scale-free prices falls to 60 percent. We also find that foreign exporters absorbed around 60 percent of the tariff burden, rejecting perfect competition and indicating reduced markups.

*We thank Nuno Limao, Charly Porcher, Daniel Xu, Justin Pierce, Davin Chor, Esteban Rossi-Hansberg, Frederic Warzynski, and Xiang Ding for early conversations and Nahim Bin Zahur for discussion. Any views expressed are those of the authors and not those of the US Census Bureau, the Board of Governors of the Federal Reserve System, or any other person associated with the Federal Reserve System. 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 (CBDRB-FY25-P2193-R12042, R12792). Data analysis and coding was entirely done by humans, and some text editing was done using AI tools.

1 Introduction

Determining whether foreign producers or domestic buyers ultimately bear the burden or incidence of tariffs is a central question in international economics. A critical element in determining incidence is the pass-through of tariffs into prices (Jenkin, 1872). In this paper, we revisit tariff incidence and pass-through in the 2018–2019 period, taking into account an underappreciated feature of US trade: transaction-level quantity discounts. Using comprehensive US microdata, we show that these discounts account for the majority of price variation within narrowly defined product categories, and are inconsistent with the typical interpretation of the 2018–2019 period. Our findings reveal that accounting for quantity discounts reduces measured tariff pass-through, with foreign exporters shouldering a much larger share of the tariff burden than previously recognized.

In our empirical analysis, we distinguish between factors that affect the level of prices (such as firm productivity or market power) and those that affect the scale of prices (how unit prices change with transaction size).[1] If scale reflects price discrimination, tariffs may simply transfer rents. If scale reflects real resource costs (e.g., logistics), shrinking order sizes represent a rise in real unit costs - a loss of resources. While market power may shift price levels, we find that it does not explain the steepness of the quantity discount schedule. Specifically, we estimate that a 1% increase in transaction size leads to a 0.29% reduction in price—a "scale effect" that is robust across relationships and driven by genuine supply-side economies rather than price discrimination.

We aggregate the transaction-level data into unit values and run a series of tariff pass-through regressions. We find that tariff pass-through to post-tariff buyer prices, as estimated in prominent work, is about 100 percent. However, this result masks two important underlying forces. First, we find that these measures of pass-through (near 100 percent) conflate two opposing forces: a strategic price cut by exporters and a mechanical price rise due to the loss of scale economies. We estimate that the 'scale-free', strategic pass-through—reflecting the shift in the exporter's price schedule—is only 60 percent. Second, had exporters not cut markups, our results imply that observed tariff pass-through would have exceeded 100 percent given the downward sloping supply curve. While we do not account for any foreign retaliation, this provides a micro-founded baseline for future research.[2]

A striking feature of both domestic and international markets is the substantial price variation observed across seemingly identical products—a fact well documented in the

[1] We define the "scale-free price" ($\tilde{p}$) as the price of a single unit, netting out volume discounts.

[2] Extending to China's retaliatory tariffs and conducting general equilibrium analysis are beyond our scope.

literature (e.g., Stigler, 1961; Halpern and Koren, 2007; Atalay, 2014; Fontaine, Martin, and Mejean, 2020; Bornstein and Peter, 2025; Burstein, Cravino, and Rojas, 2024). Bornstein and Peter (2025) and Meleshchuk (2017) interpret these price differences entirely through the lens of markup differences across orders of different transaction sizes. Similarly, Alviarez, Fioretti, Kikkawa, and Morlacco (2023) assume market power and Burstein et al. (2024) assume exogenous wedges. We find little empirical support for these mechanisms in explaining quantity discounts in US trade. Instead, our results align with the distribution and operations research literature (Munson and Rosenblatt (1998), Davis, Grim, Haltiwanger, and Streitwieser (2013), Munson, Jackson et al. (2015), Hornok and Koren (2015)), as we find that quantity discounts are driven by scale economies and fixed costs that must stem from the underlying production and demand processes.[3] Our analysis demonstrates that these discounts are not merely anecdotal but are a pervasive and quantitatively important feature of US trade.

The implications for tariff pass-through are large. Estimates of the incidence of the 2018–2019 tariffs are that they fell almost entirely on US buyers at the border, with pass-through rates nearly unity (Amiti, Redding, and Weinstein, 2019; Gopinath, Itskhoki, and Rigobon, 2010; Cavallo, Gopinath, Neiman, and Tang, 2021; Fajgelbaum, Goldberg, Kennedy, and Khandelwal, 2020). To explain this result in a standard model setup requires that exporting firms have constant markups and that the export supply curve facing the US is perfectly elastic (resulting in a flat marginal costs curve). In that setup, exporters bear none of the tariff burden. In contrast, we estimate that the export supply curve facing the US slopes downward (as is needed to explain quantity discounts).

Our approach demonstrates that the way scale economies aggregate across transactions is fundamental to understanding overall tariff pass-through. By using detailed transaction-level data, we show that micro-level quantity discounts and scale economies, when aggregated, directly determine the observed market-level incidence of tariffs. This perspective moves beyond traditional models that assume uniform pricing or infer scale economies from aggregate data (e.g., Antweiler and Trefler (2002); Bartelme, Costinot, Donaldson, and Rodriguez-Clare (2025); Lashkaripour and Lugovskyy (2023)). We extend the standard pass-through framework (e.g., Weyl and Fabinger (2013); Amiti et al. (2019); Ganapati, Shapiro, and Walker (2020)) by showing that aggregate pass-through is not a primitive, but an outcome shaped by the distribution of transaction sizes and the prevalence of non-linear pricing. This approach clarifies how micro-level pricing behav-

[3]This is complementary to models that emphasize transaction costs and search frictions (e.g., Allen (2014); Krolikowski and McCallum (2025)), but we do not explicitly model these frictions. Instead, we capture their effects to the extent that they manifest as scale economies or quantity discounts in observed prices.

ior translates into macro-level effects.

Our single correction helps resolve three puzzles in international trade. First is the pass-through puzzle: recent empirical work finds complete tariff pass-through to US import prices (Fajgelbaum et al., 2020; Cavallo et al., 2021), a result that in standard models implies foreign export supply is perfectly elastic. This contradicts the consensus that the US is a large open economy, and the views of the optimal tariff literature (Broda, Limao, and Weinstein, 2008; Ossa, 2014) that US tariff pass-through should be incomplete. We estimate that supply is actually downward-sloping ($\gamma < 0$); thus pass-through only *appears* complete because plummeting transaction volumes push unit costs up the supply curve, masking the strategic price cuts by exporters.

Second is the incidence puzzle: while US import prices suggest domestic consumers bore the entire tariff burden, independent evidence from foreign production sites reveals significant welfare losses for exporters (Chor and Li, 2024). Indeed, typical model parameters imply that 70 percent of tariff incidence would fall on the foreign exporter (see Miran (2025)). This creates a tension: if exporters fully passed on the tariffs, how could they suffer losses? We resolve this by showing that strategic pass-through was incomplete. Exporters lowered their effective markups and prices through quantity adjustments, thereby absorbing a significant share of the incidence, consistent with the observed contraction in foreign activity. We estimate that around 60 percent of tariff incidence fell on foreign exporters, rejecting perfect competition.

Third is the exchange rate disconnect puzzle: while tariff prices appear to move one-for-one, exchange rate pass-through is notoriously incomplete, typically around 0.5 (Gopinath et al., 2010; Amiti, Itskhoki, and Konings, 2014). By correcting unit values for quantity distortions, we align tariff pass-through estimates (≈ 0.6) with these established exchange rate findings, suggesting a unified pricing behavior across shocks.

The rest of the paper proceeds as follows. We first decompose transaction-level prices and show that the size of transaction explains a very large portion of the dispersion in observed prices. Second, we show this is a largely supply-side phenomenon, where shifts in demand identify the transaction-level supply elasticity. Third, we show that variation in quantity discounts is driven by observable cost factors (shipping costs and within-firm transactions), as well as being higher in markets with greater price dispersion (a common proxy for higher search frictions). Fourth, we consider aggregation and show that reduced-form estimates of pass-through require important additional interpretation, given quantity discounts. Finally, we revisit the question of tariff incidence.

2 Conceptual Framework

Analyzing tariffs with quantity discounts requires a framework accommodating flexible supply slopes and potentially non-linear pricing (Stole, 2007). Classic analysis typically assumes upward-sloping supply and uniform pricing (Weyl and Fabinger, 2013). We review the uniform pricing benchmark before formalizing our transaction-level decomposition. This approach sets the stage for our later empirical analysis, which assesses the prevalence and consequences of quantity discounts in US import data.[4]

Under uniform pricing and specific taxes, pass-through ($\rho = \partial p/\partial t$) is well-defined.[5] Incidence is the ratio of downstream to producer welfare changes: $I \equiv (\partial DS/\partial t)/(\partial PS/\partial t)$. Under monopoly, $I = \rho$; under perfect competition, $I = \rho/(1-\rho)$. Conditional on market structure, pass-through is a sufficient statistic for local incidence.[6]

Non-uniform pricing complicates this logic. Assume a continuum of customers optimally choosing quantities.[7] If a monopolist implements first-degree price discrimination, they price along the demand curve, capturing all surplus. With a tariff, a single pass-through figure is undefined, as each customer type faces a different price and pass-through rate. Aggregate pass-through is uninformative. In this case, incidence is mechanically 0 ($I = 0$), as all surplus loss falls on the producer. Conversely, under perfect monopsony where the buyer extracts all surplus, a tariff extracts all buyer surplus, and incidence is infinite ($I \to \infty$).

We decompose the price of transaction t between seller i and buyer j for variety v and quantity q with tariffs τ as:

$$\log P_t = \log \tilde{p}_t + \log p(q_t) = \log \tilde{c}_t + \log \tilde{\mu}_t + \log c(q_t) + \log \mu(q_t) + \log(1+\tau_t). \quad (1)$$

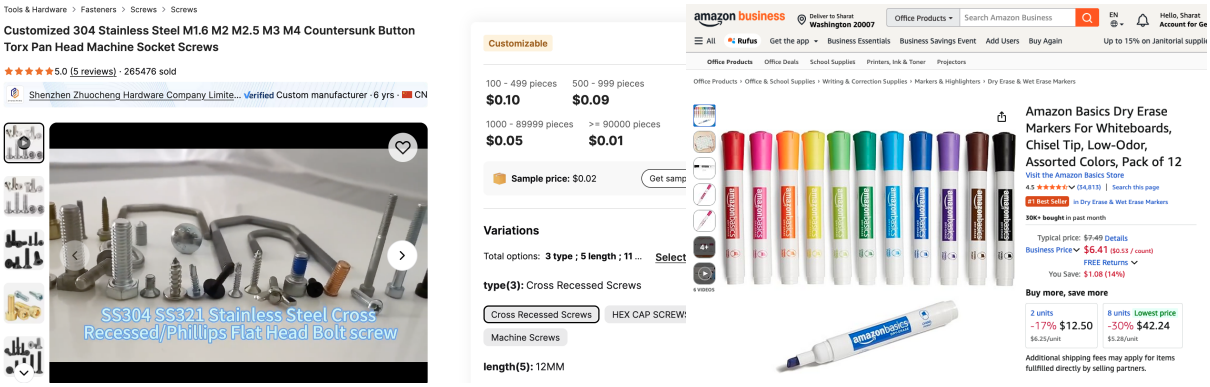
Normalizing $c(1) = \mu(1) = 1$, level terms $\tilde{p} = \tilde{c}\tilde{\mu}(1+\tau)$ capture product- and firm-level heterogeneity, while scale terms $p(q) = c(q)\mu(q)$ reflect variation with size. This distinguishes scale economies from level effects. Since costs and markups vary with quantity, a single price is undefined. However, we can still analyze the local effects of a policy by

[4] While these results have been explored theoretically (Maskin and Riley, 1984), empirical study is largely neglected. (Exceptions include Meleshchuk (2017) in trade and Verboven (2002) in industrial organization).

[5] For exposition, we use a specific tax. An ad-valorem tax, such as a tariff, changes the derivations, but the same logic applies (Bishop, 1968; Cheung, 1998).

[6] The relationship between incidence and pass-through depends on the assumed market structure. While $I = \rho$ under monopoly and $I = \rho/(1-\rho)$ under perfect competition, intermediate cases of oligopoly can also be characterized. For example, under symmetric Cournot competition, incidence can be expressed as a function of both the pass-through rate and a "conduct parameter" θ that captures the degree of market power, as formalized in the general framework of Weyl and Fabinger (2013).

[7] We focus on quantity, but the framework is isomorphic to quality (Melitz, 2003).



(a) Screw Bolts on Alibaba

(b) Markers on Amazon Business

Figure 1: Real World Examples of Quantity Discounts

aggregating up from the transaction level.[8]

Figure 1 shows real-world examples of quantity discounts from Alibaba, a large online marketplace and Amazon Business. In both panels, we see that as the quantity purchased increases, the unit price decreases significantly. In the first example (Figure 1a), we would say that $\tilde{p} = 0.10$ and $p(q)$ decreases from 0.10 to 0.04 as quantity increases from 1 to 100. The elasticity of $p(q)$ to q is approximately -0.35.

2.1 Framework: Aggregate Scale Economies

Equation (1) defines transaction-level scale economies. We also consider how aggregate prices and costs respond to shocks. Using the scale-free price $\tilde{p}_t = \tilde{\mu}_t \tilde{c}_t$, we examine how an aggregate price index $\mathbf{P}$ changes with non-linear pricing.

We define the aggregate price index $\mathbf{P} = \sum_{t \in T} w_t \tilde{p}_t$, where $\tilde{p}_t = P_t / p(q_t)$ is the scale-free price and w_t are weights. This separates cost/markup changes from compositional effects, decomposing aggregate scale economies $\mathcal{S}_P$ into level effects ($\tilde{p}$) and scale effects ($p(q)$).

2.2 Framework: Tariff Incidence

We analyze the local incidence of tariffs on downstream buyer surplus, producer profits, and government revenue. Buyer surplus changes are derived by integrating under the demand curve (Weyl and Fabinger, 2013):

$$DS(\mathbf{p}) \equiv \int_{t \in ijvt} \int_{P_t}^{\infty} D(z) dz dt,$$

[8] We assume scale terms are common across transactions; we test this in the empirical section.

such that to a local approximation, defining a scale-free pass-through as $\tilde{\rho} \equiv \frac{d\log \tilde{p}_t}{d\log(\tau+1)}$,

$$\frac{dDS}{d(\tau+1)} = -\int_t \tilde{\rho}\frac{E_t}{1+\tau}dt \tag{2}$$

where E_t represents expenditures.

This computation does not require demand curvature, which is only needed for global results and follows from individual buyers internalizing how changes in orders affect their prices.

Suppose we use unit values where $\rho \equiv \frac{d\log P_t}{d\log(1+\tau)}$ as a reduced-form pass-through. Then, pass-through estimated using unit values conflates strategic price changes with mechanical scale effects:

$$\tilde{\rho} = \rho - \frac{d\log p(q_t)}{d\log q_t}\frac{d\log q_t}{d\log(1+\tau)}.$$

If scale economies are large and quantities fall with tariffs, pass-through will be overestimated. Conversely, if we use ρ instead of the true pass-through of scale-free prices $\tilde{\rho}$, the change in downstream surplus will be overestimated.[9] However, the unit pass-through ρ still plays a role, as it is linked to producer surplus, which we now turn to.

Producer surplus is defined as:

$$PS \equiv \int_{t\in ijvt}[R(t) - C(t)]\,dt,$$

where $R(t)$ is the revenue from a transaction and $C(t)$ are the costs of a transaction.

We can then differentiate with respect to the tariff rate:

$$\frac{dPS}{d(\tau+1)} = \int_t \left[\frac{dR(t)}{d(\tau+1)} - \frac{dC(t)}{d(\tau+1)}\right]dt. \tag{3}$$

Revenue is observable; costs are not. Using Equation (1), we decompose revenue into costs and markups. We require: (1) scale economies and price discrimination ($c(q), \mu(q)$); (2) the response of revenue and quantity to tariffs; and (3) base costs $\tilde{c}$ or markups $\tilde{\mu}$. We estimate the first two empirically and calibrate the third from literature.

[9] Appendix A.1.1 allows for $p(q_t)$ to directly vary with $1+\tau$.

Revenue changes can be expressed in terms of local elasticities:[10]

$$\frac{dR}{d\left(1+\tau\right)}=\left[\frac{d\log\tilde{p}}{d\log\left(1+\tau\right)}+\left(1+\frac{d\log p\left(q\right)}{d\log q}\right)\frac{d\log q}{d\log\left(1+\tau\right)}-1\right]\frac{R}{1+\tau}.$$

Cost changes are similarly decomposed:

$$\frac{dC}{d\left(1+\tau\right)}=\left[\frac{d\log\tilde{c}}{d\log\left(1+\tau\right)}+\left(1+\frac{d\log c\left(q\right)}{d\log q}\right)\frac{d\log q}{d\log\left(1+\tau\right)}\right]\frac{C}{1+\tau}$$

We decompose the price-quantity elasticity into scale economies $\gamma_c \equiv \frac{d\log c(q)}{d\log q}$ and price discrimination $\gamma_\mu \equiv \frac{d\log \mu(q)}{d\log q}$.

We use the 2018–2019 US tariffs to identify changes in scale-free prices, quantities, and costs. Recovered scale economies allow us to decompose price changes and compute incidence and surplus changes.

Government revenue is: $G \equiv \int_t \tau R(t)$. The change in government revenue with respect to the tariff rate is:

$$\frac{dG}{d\left(1+\tau\right)}=\int_t\left[\tau\frac{dR\left(t\right)}{d\left(1+\tau\right)}+R(t)\right]dt.$$

Global incidence requires estimating supply and demand curves and integrating, necessitating functional form assumptions.

Before we can implement the framework above, we first turn to a basic question. What is the nature of price variation in the underlying data, and as decomposed by equation (1)? Is there even a correlational link between transaction size and price?

3 Decomposing Prices

3.1 Data

We focus on the universe of US imports from abroad. 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. This combined database, the Longitudinal Firm Trade Transactions Database (LFTTD), is the source for every shipment transaction entering the US via air, water, or land.

[10] We define prices as the tariff-inclusive price (FOB price plus tariffs). We revisit the difference with the freight inclusive price (CIF) in Section 5.4. Additionally, if we know market conduct, the incidence ratio $I = \frac{dCS/d\log T}{dPS/d\log T}$ is: $I = \frac{-\text{Expenditure}\times\tilde{\rho}}{-\text{Revenue}\times[1-(1-\theta)\rho]} \approx \frac{\tilde{\rho}}{1-(1-\theta)\rho}$

The data consider exporters (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 (Kamal and Monarch, 2018). These downstream US firms are a mix of manufacturers, wholesalers, and retailers. Our units of observation are individual transactions of an exporting firm, importing firm, 10-digit US Harmonized Tariff Schedule code (HTS-10), port of entry, and date of entry. For each transaction, we observe the value, shipping and freight charges, method of transport, tariffs and customs charges, a standardized measure of quantity within that HTS code, and the weight of the shipment. Of note, quantities and weights are often equivalent for many HTS codes, as goods are often sold by weight.

Prices are computed as value/quantity, though further analysis will also compute prices inclusive of tariff and freight charges. We decompose prices for 2017 and report robustness using 2012 data. In 2017, 50% of trade transactions occurred between parties that trade more than 100 times within an HTS-10 code in a year.

Alternative Data For domestic trade, we consider the US Census and Department of Transportation Commodity Flow Survey (CFS), focusing on 2012. See Appendix C.1 for details. The downside of this data is the lack of information on the identity of the buyer. Collectively these two datasets cover the entire universe of final and intermediate physical goods consumed in the United States.

3.2 Decomposing Import Prices

We demean log variables by variety (HTS-10). [11] We decompose prices into a scale component (γ) and a level component (μ). The relationship for transaction t is:

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

where γ is the quantity scale elasticity and $\mu_{i\to j,v}$ is the relationship-specific price level. μ absorbs time-invariant factors (markups, quality, bargaining), isolating the scale effect γ. Index v is HTS-10 and source country; i is seller; j is buyer.[12]

Table 1 decomposes 2017 import prices for 60 million residualized transactions. A uniform quantity relationship explains 44.6% of price variation (Spec 1). Allowing quantity discounts to vary by HTS 6-digit (following the logic of Rauch (1999)) code explains 57.6%

[11] We also explored time variation, demeaning by HTS-10 and month.

[12] Samples are aligned to include only observations present in all decompositions, removing singletons (dropping about 5% of observations).

of variation (Spec 2).

A stricter test looks within Buyer-Seller-Variety relationships. Seller-Variety fixed effects account for 73% of the price variation in the data. Buyer-Variety fixed effects account for 50% of the price variation. These highly disaggregated controls absorb 73.3% of the variation in prices (Specification 3). Adding a uniform quantity discount explains 38.3% of the variation (specification 4; 17.4% directly and 20.9% through the covariance). Varying discounts by HTS 6-digit strengthens this (Spec 5). Even with millions of relationship fixed effects, transaction size explains a large share of price variation.

We compare this to yearly aggregate measures (market power, total annual volume) in Specifications 6-10.

The aggregate quantity (within variety), bought by a particular buyer, sold by a particular seller, and sold from a particular seller to buyer, is defined:

$$\log\left(q_{i,v}\right) \equiv \log \sum_{i',v'=i,v} q_{i'\to j,t,v'}, \qquad \log\left(q_{j,v}\right) \equiv \log \sum_{j',v'=j,v} q_{i\to j',t,v'},$$
$$\log\left(q_{i,j,v}\right) \equiv \log \sum_{i',j',v'=i,j,v} q_{i'\to j',t,v'}. \tag{5}$$

Aggregate volume measures explain only 6.5% of variation (Spec 6), compared to 57.6% for transaction-level quantity (Spec 2). Adding aggregate measures to transaction-level quantity adds minimal explanatory power, increasing the R-squared from 44.6% (Spec 1) to just 45.0% (Spec 7). This suggests that discounts are driven by per-shipment scale economies — such as in logistics, production, or handling — rather than by bargaining power derived from aggregate purchasing volume.

We measure traditional market power using bilateral market shares (Bernard, Dhyne, Magerman, Manova, and Moxnes, 2022):

$$s_{i\to j,v} = \frac{q_{i,j,v}}{q_{i,v}}, \qquad s_{i\leftarrow j,v} = \frac{q_{i,j,v}}{q_{j,v}}. \tag{6}$$

Shares account for only 0.8% of variation (Spec 8). Specification 9 confirms minimal additional explanatory power. Compared to seller/buyer fixed effects (Spec 10, 11), relation-level aggregates have minimal power.

The key takeaway from Table 1 is the dominance of specific transaction size over aggregate leverage. Standard theories of bargaining power or second-degree price discrimination suggest that total volume (Specification 7) or market share (Specification 8) should drive discounts. Instead, we find that the specific size of a single shipment is the primary explanatory factor. This suggests a logistics and supply-chain interpretation, where fixed

Table 1: Correlating Price Variation: Decomposition of Log(Price) Variation

	Specification		Variance Decomposition			
	Controls	Fixed Effects	Controls	Fixed Effects	Covariance	Residual
1	$\log q$		44.6%			55.4%
2	$\log q$ x HTS6		57.6%			42.4%
3		Seller-Buyer-Variety		73.3%		26.7%
4	$\log q$	Seller-Buyer-Variety	17.4%	41.6%	20.9%	20.1%
5	$\log q$ x HTS6	Seller-Buyer-Variety	31.5%	32.0%	20.2%	16.3%
6	Aggregate Quantities		6.5%			93.5%
7	+ $\log q$		45.0%			55.0%
8	Relationship Shares		0.8%			99.2%
9	+ $\log q$		44.6%			55.4%
10		Seller-Variety, Buyer-Variety		72.5%		27.5%
11	Relationship Shares	Seller-Variety, Buyer-Variety	1.5%	72.6%	-1.5%	27.4%

Notes: This table decomposes 2017 import transaction-level price variation after demeaning log prices and quantities by country-origin and variety (HTS-10) fixed effects. Sellers are designated at the Manufacturer ID and variety. Buyers are designated at the domestic firm ID and variety. For consistency, the sample is fixed to remain constant over all specifications. See the text for full details and sources.

costs of packaging, shipping, and handling create economies of scale at the batch level, over a power-based interpretation where "important" buyers extract consistently lower prices for a particular transaction. Even controlling for the identity of the buyer and seller (Specification 4), which absorbs permanent bargaining power, the transaction-level quantity effect remains robust.

Alternative Sample: Domestic Trade Data In domestic trade data we broadly find similar trends to the international trade import data. Appendix Table A.8 replicates the exercise. Within tightly defined varieties, a simple log-linear quantity discount explains between 30-40% of all price variation. See Appendix C.1.1 for further details.

These descriptive regressions show transaction size predicts price, even with fixed effects. Scale effects correlate to over half of price variation, distinct from level effects. We next establish causality, isolating supply-side scale economies from demand effects.

4 Scale Economies and Quantity Discounts in Purchases

Is the price-quantity link driven by supply-side scale economies or downward-sloping demand? We isolate the supply-side to recover the buyer's price schedule. Identifying the transaction-level supply slope is challenging. Negative price-quantity correlation could reflect scale economies or a demand curve.

We adopt a log-linear specification, a choice naturally fitting the local incidence framework. We decompose the price of a transaction t between seller i and buyer j for variety v and quantity q as follows:

$$p_{i\to j,t,v} = \tilde{p}_{i\to j,v} q_{i\to j,t,v}^{\gamma} \xi_{i\to j,t,v}. \tag{7}$$

$\tilde{p} = \tilde{\mu}\tilde{c}$ is the unit price shifter. ξ captures unobserved quality/measurement error. $\gamma = \gamma^c + \gamma^\mu$ combines cost-scale (γ^c) and price-discrimination (γ^μ) elasticities.

The demand equation is:

$$q_{i\to j,t,v} = \tilde{q}_{i\to j,t,v} f\left(p_{i\to j,t,v}\right), \tag{8}$$

Following Berry and Haile (2021), we use high-frequency demand shocks $\tilde{q}_{i\to j,-t,v}$ to avoid endogeneity with unobserved quality $\xi_{i\to j,t,v}$.

In Appendix B.3, we also put explicit structure on the demand following Feenstra (1994). In this situation with high frequency data, we assume that variation in residual supply and demand across different sellers within a buyer and across time (after double differencing) are orthogonal to each other. Both results find similar results for the curvature of the pricing curve. We next detail our instrumental variable and fixed effect strategy.

4.1 Scale Pricing With Instrumental Variables and Fixed Effects

The estimating relationship is:

$$\log p_{i\to j,t,v} = \gamma \log q_{i\to j,t,v} + \mu_{i\to j,v} + \xi_{i\to j,t,v}. \tag{9}$$

The scale elasticity γ governs how quickly prices change as quantity ordered increases within a buyer-seller relationship. We use buyer-seller-variety fixed effects, identifying off within-relationship quantity variation. We micro-found downstream demand shocks affecting the optimal inventory stocking of a downstream firm (Arrow, Harris, and Marschak, 1951; Baumol, 1952), with firms ordering different levels of a variety over the course of a year. This model also produces variation in order frequencies, consistent with the lumpiness of international trade flows documented in inventory-based models (Alessandria, Kaboski, and Midrigan, 2010).

This strategy is in the spirit of Hillberry and Hummels (2013), but differs in a critical element: we consider transactions and abstract away from any measures of aggregates of

transactions or direct measures of classic market power.

Causal identification faces two threats: unobserved relationship characteristics and simultaneous supply curve shifts. Unobserved quality variation (e.g., different iPhone models under one HTS code) could bias results. If quantity varies inversely with quality, this creates a mechanical negative relationship, opposite to the standard IO bias in demand estimation. This case resembles third-degree price discrimination or Alchian-Allen effects (Hummels and Skiba, 2004). We require a time-varying demand shock at the buyer level to address these threats.

Unobserved Characteristics Are we mismeasuring quality within even the highly disaggregated HTS-10 categories and a trading pair of downstream buyers and factories? For example, a buyer may purchase 10,000 iPhone 16s at $500 and 1,000 iPhone 17s at $1000 and both under HTS code "8517.14.0050". If goods are sold in efficiency units, we need a quantity shifter, uncorrelated with quality ξ.

This would be consistent with some degree of third-degree price discrimination. It also would be consistent with a form of the Alchian–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.

Measurement error A related issue is measurement error in the independent variable, quantity.[13] Measurement error has two potential links: first is in q itself, and second in its implicit link to p through the fact it is generated using q. The first can mechanically induce attenuation bias if q is observed with classic measurement error. The second also induces bias by creating a mechanical negative correlation between the error in the derived price and the mismeasured quantity q. As with standard measurement error, an instrument solves the additional error implicit in p through mismeasured q. We now turn to an instrumentation strategy.[14]

Instrumental Strategy Instrument validity depends on $Var(q, \xi)$. Positive covariance (larger quantities of higher quality) biases OLS upward; negative covariance biases it

[13] We assume that transaction values are reported without error, as tariffs are charged on them and are closely monitored.

[14] Our unit of observation is a transaction, thus relationships with more transactions are more influential, echoing the weight correction term in estimators such as Broda and Weinstein (2006).

Table 2: Recovering Quantity Discounts: OLS Results

	(1)	(2)	(3)	(4)	(5)
			$\log p$		
$\log q$	-0.83	-0.533	-0.469	-0.296	-0.287
	(0.0127)	(0.00249)	(0.00275)	(0.00328)	(0.00354)
R^2	0.766	0.913	0.446	0.797	0.81
Within R^2	0.766	0.531	0.446	0.253	0.244
Fixed Effects					
Variety		✓			
Buyer-Seller-Variety				✓	✓
Seller-Month-Variety					✓
Demeaned					
Month-Variety			✓	✓	✓
Country-Variety			✓	✓	✓

Notes: Round parentheses represent standard errors clustered at the relationship level. Demeaning regularizes all variables by country-origin and product variety (HTS-10) fixed effects, as well as month and variety fixed effects. Sellers are designated at the Manufacturer ID and variety level. Buyers are designated at the domestic firm and variety level. Standard errors are clustered at the relationship. See the text for full details and sources.

downward. We require a high-frequency instrument. Assuming goods within an HTS-10-buyer pair are linked through the buyer's problem, changes in other purchases act as instruments.

With our high degree of fixed effects, we need a fine-grained instrument for $q_{i\to j,t,v}$ that varies at the monthly or weekly transaction level. This instrumental strategy relies on the demand-side variation, exogenous shifts in buyer demand that trace out the supply curve.

We instrument transaction size $q_{i\to j,t,v}$ with the total quantity of variety v purchased by buyer j from other transactions in the same month:

$$IV_{i\to j,t,v} \equiv \log \sum_{t\neq t', t'\in(month_t)} q_{i\to j,t',v}. \tag{10}$$

This leverages buyer-level demand shocks, assumed orthogonal to seller-specific quality shocks.

The validity of this instrument rests on the assumption that buyer-level demand shocks for a variety v are correlated across purchases, while being orthogonal to seller-specific supply or quality shocks $\xi_{i\to j,t,v}$. We use exogenous shocks to a buyer's total demand for a

variety to trace out the supply curve for specific transactions. For example, a downstream demand shock for a buyer's final product would increase its demand for input v from all its suppliers, but would be uncorrelated with a temporary quality issue from a single supplier i. Because the instrument proxies for the buyer's total downstream activity, it shifts demand for inputs. Whether inputs are substitutes or complements determines the shape of the demand curve, but the shift (the instrument) remains correlated with quantity demanded from all suppliers, orthogonal to supplier-specific supply shocks.

This strategy exploits variation in buyer procurement and inventory (formalized in Appendix A.5). Buyers face monthly demand shocks $\varepsilon_{j,t}$ that determine total procurement, allocated across suppliers based on quantity discounts. Each supplier's monthly quantity is then split across transactions for logistical reasons—shipping schedules, inventory timing, payment terms.

To control for time-varying supply shocks, we include another set of fixed effects.

Supply Shocks Is identifying variation coming off a shift in the supply curve, rather than a shift in demand? To ensure we identify the supply curve, we control for temporal supply shocks as many trading relationships are long-lived: $\gamma_{i,v,t\in M_t}$, where M_t is the month of transaction t. These stringent fixed effects reduce sample size, dropping 30% of value and transactions.[15]

The inclusion of seller-month-variety fixed effects is critical for identification. By comparing the same seller across different buyers in the same month, we absorb any time-varying supply shocks—such as excess inventory or seasonal cost changes—that might be correlated with buyer demand.

A threat to identification would be non-linear pricing based on total monthly volume (rebates). However, as shown in Specification 7 of Table 1, aggregate volume has negligible explanatory power.

We also conduct robustness checks by varying the construction of the instrument, such as using alternative time windows (weekly, quarterly) and looking only within a relationship. We also exploit a slightly different strategy that constructs an IV excluding all purchases from the same supplier. This captures cross-supplier correlation from common demand shocks. Exclusion holds if supplier-specific shocks don't affect total demand or other suppliers' allocations.[16]

[15] As robustness exercises, we controlled for time-varying country-level characteristics ($\gamma_{c,v,t}$), relationship length within a year, and only arm's-length relationships.

[16] We also conduct robustness on products that are less likely to be measured with error - those that are measured in kilograms only. We also only consider q_t that is small to aggregate q bought in that period. In the next section, we interact γ with a host of buyer and supplier characteristics, we also show that base

Table 3: Recovering Quantity Discounts: IV Results

	(1)	(2)	(3)	(4)	(5)
	OLS	IV Estimates: log p			
log q	-0.268	-0.284	-0.315	-0.288	-0.200
	(0.00463)	(0.00151)	(0.00239)	(0.00152)	(0.0460)
First Stage Coeff.		-0.773	-0.654		0.011
		(0.00446)	(0.00500)		(0.00103)
First Stage F-Stat		30040	8425	10020	115
R^2	0.800				
Within R^2	0.226	0.226	0.219	0.225	0.218
IV		Baseline	Within-Relationship	3+4	Other Suppliers
Fixed Effects					
Buyer-Seller-Variety	✓	✓	✓	✓	✓
Seller-Month-Variety	✓	✓	✓	✓	
Demeaned					
Month-Variety	✓	✓	✓	✓	✓
Country-Variety	✓	✓	✓	✓	✓

Notes: Round parentheses represent standard errors clustered at the relationship level. The sample for column (1) replicates the OLS specification for this sample. Demeaning regularizes all variables by country-origin and variety (HTS-10) fixed effects, as well as month and variety fixed effects. Sellers are designated at the Manufacturer ID and variety level. Buyers are designated at the domestic firm and variety level. The J-statistic for overidentification is 5.722. The first-stage coefficients reflect the distinct mechanisms underlying each instrument: negative for within-month allocation (Columns 2 and 3, leave-one-out structure) and positive for cross-supplier variation (Column 5, common demand shocks). See Appendix A.5 for the complete microfoundation. Standard errors are clustered at the relationship. See the text for full details and sources.

4.2 Transaction-level Supply: Import Regression Results

We present regression results for the scale elasticity γ (Equation (9)). Table 2 shows OLS results. The base regression shows -0.830 (Col 1), but clearly mixes supply and demand. Simple demeaning yields -0.469 (Col 3). Adding buyer-seller-variety fixed effects reduces this to -0.296 (Col 4). Including seller-month-variety fixed effects yields -0.287 (Col 5). Thus, a 10% transaction size increase lowers price by 2.9%.

Table 3 presents IV results using the instrument from Equation (10). The first stage is strong (Col 2). The second-stage estimate is -0.284, close to OLS (Col 1). A related instrument (constructed using only the total other purchases from the same manufacturer within the relationship) yields similar results (Col 3), with $\gamma \approx -0.315$, with a first stage coefficient of -0.654. Using both instruments finds similar results (Col 4).

prices are uncorrelated with changes in total volume. We can also use a placebo test with future purchases, which should be uncorrelated with current transaction quantity.

Column 5 provides an alternative identification strategy. Rather than exploiting within-month transaction allocation (Columns 2-4), it uses cross-supplier variation from common demand shocks. The weaker first stage (F=115) reflects reduced sample size (requires multiple suppliers per buyer-month) and noisier variation (cross-supplier correlation weaker than within-month mechanical constraint).

Our two IV strategies exhibit different first-stage signs, reflecting the distinct economic mechanisms at work. The baseline specification (Column 2) shows a negative first-stage coefficient (-0.773), while the cross-supplier specification (Column 5) exhibits a positive coefficient (0.011).

These contrasting signs are expected (Appendix A.5). The negative coefficient arises from the leave-one-out structure of within-month allocation. When buyers split monthly procurement totals across transactions, larger allocations to one transaction mechanically reduce allocations to others - creating negative correlation. This is standard in peer effects literature (Angrist, 2014) where group-level constraints induce negative within-group correlations. The positive coefficient in Column 5 reflects common demand shocks across suppliers: when a buyer faces increased downstream demand, purchases from all suppliers rise together.

Both instruments satisfy the exclusion restriction under different assumptions. For Column 2, transaction-specific shocks must be orthogonal to monthly procurement plans (such as predetermined by production schedules) and allocation timing (like those determined by logistics). For Column 5, supplier-specific shocks must not affect total buyer demand or allocations to other suppliers.

Table 4 summarizes heterogeneity across HTS 6-digit categories. We re-estimate our preferred IV specification separately for each category to recover γ_v. In both the full OLS (Col 3) and IV specifications (Col 5), most products exhibit significant discounts, with mean (-0.275, -0.284) and median (-0.223, -0.239) elasticities consistent with the aggregate. We use Column (5) as a baseline in our incidence calculations.

Domestic Data We estimate supply using domestic data from the CFS in Appendix C.2. Using a similar IV strategy, we find γ_v between -0.21 and -0.26. While slightly smaller in magnitude, these estimates are consistent with our international trade results, suggesting similar supply-side scale economies in domestic and international transactions.

Structural Estimation For robustness, we employ a structural approach to jointly estimate supply and demand elasticities, following Feenstra (1994). This method relies on the heteroskedasticity of orthogonal supply and demand shocks for identification rather

Table 4: Heterogeneity in Quantity Discounts: Variation across Products

		(1)	(2)	(3)	(4)	(5)
				$\log p$		
$\log q$	median	-0.344	-0.267	-0.222	-0.241	-0.239
$\times$ Variety	mean	-0.367	-0.305	-0.275	-0.310	-0.284
	variance	0.211	0.197	0.217	1.160	0.245
R^2	mean	0.576	0.835	0.85		
First Stage F	mean				5.05	45.3
Fixed Effects						
Buyer-Seller-Product			✓	✓	✓	✓
Seller-Month-Product				✓		✓
Demeaned						
Month-Product		✓	✓	✓	✓	✓

Notes: This table summarizes the results from various specifications estimating the price elasticity with respect to quantity. Each column corresponds to a different model specification as detailed by the fixed effects and demeaning procedures applied. See the text for full details and sources.

than traditional instrumental variables. In Appendix B.3, we aggregate transactions to the monthly level for 6-digit HS categories and assume constant elasticities of substitution and supply. The structural estimates yield a median supply elasticity (ω) of approximately -0.23 to -0.37, which closely matches our reduced-form estimate of $\gamma \approx -0.29$. Additionally, we recover a median demand elasticity (σ) of approximately 3.4. The consistency between these structural estimates and our baseline results reinforces the validity of our supply-side identification. We note that while the structural approach provides valuable cross-validation, the assumed CES demand and monopolistic competition market structure are inconsistent with our later findings on pass-through. We refer readers to Ganapati and Hottman (2026) for a consistent treatment.

Validation of Approaches In Appendix B.1, we validate our baseline estimates by comparing the variety-level scale elasticities (γ_v) recovered from our OLS and IV strategies against those obtained from alternative methods. We find a strong positive correlation between our preferred estimates and those derived from a structural model of supply and demand (following Feenstra (1994)), as well as estimates based purely on observable shipping costs. The fact that shipping costs—a direct measure of logistics technology—exhibit scale economies that are highly correlated with our overall price elasticities reinforces our interpretation that the observed quantity discounts are driven by real resource costs

rather than demand-side factors or model-specific assumptions. Our elasticities are lower than Bornstein and Peter (2025) who find $\gamma \approx -0.6$ for US consumers in the retail context, and similar to Meleshchuk (2017) who finds $\gamma \in (-0.2, -0.4)$ in Colombian trade data, though interpretation (scale vs. markups) differs.

5 The Mechanism Behind Quantity Discounts

Having established that the supply curve facing US buyers is downward sloping ($\gamma < 0$), we now determine its source. If γ reflects price discrimination ($\gamma^{\mu} < 0$), tariffs may simply transfer rents. If γ reflects real resource costs ($\gamma^{c} < 0$), shrinking order sizes represent an efficiency loss.

We distinguish between effects altering the *level* of prices (scale-free price $\tilde{p}$) and those affecting the *scale* elasticity (γ). We find the scale component is cost-driven. Relationship-level market power and bargaining explain little of variation in γ. Instead, evidence points to common, cost-based scale economies—arising from fixed transaction or shipping costs. This is supported by three facts: (1) market power proxies fail to explain heterogeneity in discounts; (2) intra-firm and arm's-length transactions exhibit identical schedules; and (3) cross-market variation correlates with cost and product attribute proxies rather than competition measures.

Recall the decomposition from Equation (1):

$$\log p(q) = \underbrace{\log \tilde{c} + \log \tilde{\mu}}_{\text{level}} + \underbrace{\log c(q) + \log \mu(q)}_{\text{scale}}.$$

Both the level terms ($\tilde{c}$, $\tilde{\mu}$) and the scale terms ($c(q)$, $\mu(q)$) can reflect a mix of cost and markup elements. The level terms capture product- and relationship-specific heterogeneity in marginal costs and markups, while the scale terms capture how costs and markups vary with transaction size.

We decompose the transaction-level elasticity $\gamma_{i \to j,v}$ into cost (γ^{c}) and markup (γ^{μ}) components:

$$\gamma_{i \to j,v} = \gamma^{c}_{i \to j,v} + \gamma^{\mu}_{i \to j,v}. \tag{11}$$

Following Stole (2007), optimal non-linear pricing links the markup component of the scale elasticity to the residual demand elasticity $\epsilon_{D,v}$ faced by the firm ($i \to j$):

$$\gamma_{v} = \gamma^{c}_{v} + (1 - 1/\epsilon_{D,v}). \tag{12}$$

This decomposition yields a clear testable prediction for both sides of the market. On the seller side, greater monopoly power (lower $|\epsilon_D|$) incentives firms to use steeper discount schedules to screen buyers and extract surplus. On the buyer side, greater monopsony power allows large buyers to demand aggressive bulk discounts ("power buyer" effects), also implying a steeper schedule. Thus, if market power drives discounts, γ should be more negative in concentrated markets. In contrast, perfectly competitive markets should simply pass through the technological cost savings γ^c.

Empirical Strategy We proceed in three steps to disentangle cost-based scale economies from markup variation:

Step 1: Common quantity discounts. We begin by imposing that all seller-buyer pairs within a variety share the same cost-based scale elasticity: $\gamma^c_{i\to j,v} = \gamma^c_v$. Under this assumption, any observed heterogeneity in $\gamma_{i\to j,v}$ across relationships must reflect differences in markup behavior $\gamma^\mu_{i\to j,v}$. We test whether variation in market power—measured by seller concentration, buyer concentration, or bilateral market shares—systematically predicts this heterogeneity. If markup variation were the primary driver, we would expect relationships characterized by greater market power to exhibit systematically different quantity discounts. Alternatively, a common γ_v across relationships could also reflect a common markup component γ^μ_v arising from symmetric market frictions or search costs faced by all buyers within a market—an interpretation we explore further in Step 3.

Step 2: Offsetting effects. If we find little explanatory power from market power measures in Step 1, a mechanical possibility remains: firms with greater market power could exhibit both steeper markup schedules with respect to quantity (higher $\gamma^\mu_{i\to j,v}$) *and* flatter cost curves (lower $|\gamma^c_v|$), such that the two effects offset and leave the total scale elasticity $\gamma_{i\to j,v}$ relatively constant. To address this, we exploit variation in vertical integration. Intra-firm transactions—where transfer prices are often set to reflect cost—provide a natural benchmark for the cost component γ^c_v. If quantity discounts are similar for within-firm and arm's-length transactions after controlling for relationship fixed effects, this suggests that the common cost component γ^c_v dominates.

Step 3: Search and transaction costs. Finally, we examine whether the common component γ^c_v itself reflects genuine production-side scale economies or instead arises from search and transaction costs on the buyer side. We test whether γ_v varies systematically with product-level characteristics that proxy for these costs. Specifically, we regress the estimated γ_v on measures of product substitutability (e.g., demand elasticity σ_v), market thickness (number of buyers and sellers), and transaction frequency.

Our findings, previewed in Tables 5 and 6, show that market power measures have

minimal explanatory power, that intra-firm and arm's-length transactions exhibit similar discounts, and that γ_v correlates most strongly with cost-side factors and product substitutability—consistent with the interpretation that $\gamma_{i\to j,v} \approx \gamma_v^c$ and $\gamma_{i\to j,v}^{\mu} \approx 0$.

Table 5: Decomposition of Price Variation by Interaction Term

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	$\log p_{i\to j,t,v}$								
Independent Variable	$\log q\times$ Related	$\log q\times \log(q_{ij})$	$\log q\times \log(q_i)$	$\log q\times \log(q_j)$	$\log q\times$ Output Share	$\log q\times$ Input Share	$\log q\times$ Rel. Length	$\log q\times$ HHI Seller	$\log q\times$ HHI Buyer
Panel A: Homogeneous Interactions									
Coefficient	-0.010	0.000522	0.000449	0.000418	0.00163	-0.0104	0.000609	-0.0169	-0.00751
	(0.00420)	(0.000195)	(0.000187)	(0.000165)	(0.00222)	(0.00402)	(0.000452)	(0.00779)	(0.00634)
Within R^2	0.0001	0.000197	0.000153	0.000157	0.0000106	0.000184	0.0000313	0.000175	0.000045
Panel B: Heterogeneous Interactions									
Mean Coef.	0.004	0.028	0.036	0.036	0.035	-0.0416	0.005	-0.202	-0.100
Median Coef.	-0.016	0.0195	0.021	0.021	-0.0226	-0.051	0.000	-0.072	-0.048
Variance Coef.	0.346	0.0833	0.0896	0.0896	1.05	0.34	0.129	0.866	1.15
99% Interval (lower)	-0.441	-0.0397	-0.0439	-0.0439	-0.707	-0.363	-0.00864	-1.01	-0.635
99% Interval (upper)	0.346	0.0833	0.0896	0.0896	1.05	0.34	0.129	0.866	1.15
Controls									
Variety Specific Quantity Discounts	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed Effects									
Buyer-Seller-Variety	✓	✓	✓	✓	✓	✓	✓	✓	✓
Seller-Month-Variety	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: This table reports regressions of transaction-level prices on log quantity interacted with various relationship characteristics, while controlling for buyer-seller-variety fixed effects and seller-month-variety fixed effects. Each column represents a separate regression where the dependent variable is the interaction of log quantity with a specific characteristic. Relationship-clustered standard errors are in parentheses. Panel A presents results from a pooled specification, while Panel B estimates separate regressions for each HTS 6-digit variety in an IV framework. The market power measures include related party status (Column 1), log relationship quantity (Columns 2-4), output and input shares (Columns 5-6), relationship length (Column 7), and HHI-based seller and buyer concentration (Columns 8-9). See the text for full details and sources.

5.1 Common Cost and Markup Structures

We first test if heterogeneity in quantity discounts $\gamma_{i\to j,v}$ correlates with market power, before we consider if there is a common markup component γ_v^{μ}. We estimate:

$$\log p_{i\to j,t,v} = \log \tilde{p}_{i\to j,v} + \gamma_v^{common} \log q_t + \beta \Gamma_{i,j,v} \log q_t + \epsilon_{i\to j,t,v}, \tag{13}$$

where $\Gamma_{i,j,v}$ represents relationship-specific deviations proxied by market power measures.

Table 5 presents the results. Panel A reports pooled specifications in the baseline framework; Panel B estimates separate regressions for each HTS 6-digit variety in the IV framework. Each column and row corresponds to a different market power measure,

including vertical integration (related party status), bilateral market shares (the share of quantity bought and sold), relationship size (total quantity purchased), relationship length, and seller/buyer concentration (HHI weights, or squared market shares).

Panel A reports the results from the pooled specification. Across all specifications, market power measures have negligible explanatory power, and often with the "wrong" sign.

Columns 3 and 4 consider the aggregate size of the seller ($\log q_i$) and buyer ($\log q_j$). We find that larger parties face slightly flatter discount schedules (positive interaction coefficients), a result that directly contradicts the view that "power buyers/sellers" leverage their absolute size to negotiate steeper marginal discounts.

Columns 5 and 6 examine bilateral dependence: the share of the seller's output purchased by the buyer (Output Share) and the share of the buyer's inputs provided by the seller (Input Share). Dependence forces might suggest that a buyer who purchases a seller's entire output (Output Share ≈ 1) or a seller who provides a buyer's entire input (Input Share ≈ 1) would negotiate different terms. However, estimating these interactions yields economically trivial coefficients (0.0016 and -0.0104 respectively). Since shares range from 0 to 1, even a fully dependent buyer faces a discount schedule nearly identical to one with negligible dependence.[17]

Finally, Columns 8 and 9 interact quantity with market concentration measures (HHI). While price discrimination theory suggests monopolists might employ steeper discount schedules to screen customers, we find coefficients of only -0.017 (Seller HHI) and -0.008 (Buyer HHI). Directionally, these signs suggest slightly steeper discounts in concentrated markets, but the magnitude is economically insignificant—shifting the elasticity by less than 0.02 from perfect competition to monopoly.

Panel B confirms this with independent regressions across HS 6-digit codes. While there is heterogeneity across products (as seen in the variance of the coefficients), median interaction effects are consistently close to zero. For example, the median coefficient for the interaction with 'Output Share' is -0.023, and for 'Input Share' is -0.051. This lack of systematic variation suggests that heterogeneous markup behavior, $\gamma^{\mu}_{i \to j,v}$, is not a significant driver of quantity discounts. However, the vast majority of these results are statistically insignificant at conventional levels. [18]

This supports the interpretation that the common component γ_v^{common} primarily re-

[17] We avoid mixing revenue-based and quantity-based shares as regressors. Since $\log p = \log(\text{revenue}) - \log(\text{quantity})$, this creates mechanical correlation between the dependent variable and regressors.

[18] Relationship size and length (Antras and Helpman, 2004) show similar patterns: larger and longer-lived relationships exhibit marginally smaller quantity discounts, but these effects remain economically modest and add minimal explanatory power beyond the baseline average quantity discount (Col 7).

flects cost-based scale economies rather than heterogeneous price discrimination across relationships.

In Appendix B.2, we go a step further and recover a unique scale elasticity for every buyer-seller pair, rather than just interacting a common elasticity with relationship characteristics. We then regress these relationship-specific elasticities on measures of market power, including bilateral volume and market shares. Consistent with the interaction results, we find no significant correlation, reinforcing the conclusion that quantity discounts are not driven by relationship-specific markup variations. While buyers and sellers with and without market power have similar scale elasticities, they may have different reasons; there may be offsetting markup and cost effects. We explore this next.

5.2 Offsetting Effects and Common Costs

Could firms with market power have steeper markup schedules (γ^{μ}) but flatter cost curves (γ^{c}), masking heterogeneity? We test this using vertical integration, where intra-firm transfer prices often reflect costs, providing a benchmark for γ_v^c.

Table 5, Column 1 shows related-party status in Customs data has a negligible effect (1 percentage point). Intra-firm and arm's-length transactions exhibit nearly identical discount schedules.[19]

Relaxing to buyer-variety and seller-variety FE confirms buyers face similar discounts from both related and non-related sellers. This supports the interpretation that γ primarily captures common cost structure γ_v^c, not heterogeneous markups $\gamma_{i\to j,v}^{\mu}$.

5.3 Market-Level Variations and Search Costs

We next examine *cross-market* patterns by regressing the estimated quantity discount γ_v on market characteristics.

If discounts were driven by market-wide markup variation (γ_v^{μ}), we would expect them to be larger in less competitive markets. To test this, we regress the estimated quantity discount γ_v on various market-level characteristics Y_v:

$$\gamma_v^{common} = \alpha Y_v + \epsilon_v, \tag{14}$$

where γ_v^{common} is the recovered quantity discount variety v which indexes a particular variety v.

[19] While transfer pricing is subject to tax incentives, identical slopes would require an improbable coincidence of tax strategies matching logistic economies.

Table 6 considers measures of market concentration (HHI) and market thickness (number of buyers/sellers or buyer-seller pairs) and results in a clear pattern: concentrated markets feature significantly shallower quantity discounts. The positive coefficients on HHI (0.193 for sellers, 0.15 for buyers in Columns 1 and 2) imply that moving from a competitive market to a monopoly/monopsony moves the elasticity γ toward zero (flattening the curve).

This result contradicts both standard monopoly and monopsony explanations for price discrimination. If powerful sellers used quantity discounts to screen buyers and extract surplus (second-degree price discrimination), we would expect steeper discount schedules in concentrated markets to disincentivize arbitrage. If powerful buyers used leverage to demand bulk discounts ("power buyer" effects), we would expect steeper discounts in markets with concentrated buyers. Instead, the steepest discounts are found in the most competitive and fragmented markets. This aligns with a cost-based view where competitive pressure forces firms to pass on the full extent of logistic scale economies (γ^c) to buyers, whereas imperfect competition may dampen this pass-through, resulting in flatter pricing schedules.

In contrast, while the standard demand substitutability proxy (demand elasticity σ) offers little explanatory power, proxies for search costs are informative. Markets with more elastic demand (higher σ) exhibit little to no change in quantity discounts in our preferred IV specification (Panel B, Column 8). This null result implies that the demand-side environment has no bearing on the discount schedule, contradicting markup-driven explanations. Instead, the discount schedule appears invariant to demand conditions, consistent with a universal cost-based mechanism where standardized goods share similar logistic scale economies regardless of substitutability.

Column (9) shows that markets with greater price dispersion, a proxy for search or transaction frictions, exhibit larger discounts. This finding points towards an alternative mechanism rooted in search and transaction costs. When buyers face fixed costs to find and vet suppliers, or when sellers face fixed costs per transaction, quantity discounts naturally arise as a way to amortize these costs over larger orders. This creates a form of scale economy, either on the buyer's side (search) or the seller's side (transaction). Such a mechanism is consistent with our finding that discounts are larger for less substitutable products, where search costs are likely to be higher.

However, a pure search cost story has trouble explaining the uniformity of discounts *within* relationships. If search costs are paid upfront to establish a relationship, we would expect the scale elasticity γ to be smaller for subsequent, within-relationship transactions. Our finding that γ is stable and significant even within long-term relationships suggests

Table 6: Quantity Discounts and Across Market Variation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	Quantity Discount γ_v								
Independent Variable	HHI Seller	HHI Buyer	log(Buyers)	log(Sellers)	log(Pairs)	log(Trans.)	log(Value)	Demand log(σ)	Variance($\tilde{P}$)
Panel A: Dependent Variable: OLS Scale Parameter by HTS-6 Code; γ_v^{OLS}									
Coefficient	0.0892	0.0904	-0.015	-0.0101	-0.00904	0.00138	0.00156	-0.0141	-0.154
	(0.0453)	(0.0378)	(0.00638)	(0.00594)	(0.00594)	(0.00500)	(0.00508)	(0.00543)	(0.0128)
R-squared	0.002	0.002	0.004	0.002	0.002	0	0	0.002	0.086
Panel B: Dependent Variable: IV Scale Parameter by HTS-6 Code; γ_v^{IV}									
Coefficient	0.193	0.15	-0.0355	-0.0283	-0.025	-0.00867	-0.0104	0.000773	-0.199
	(0.0679)	(0.0485)	(0.00687)	(0.00618)	(0.00542)	(0.00713)	(0.00716)	(0.00903)	(0.0208)
R-squared	0.001	0.001	0.003	0.002	0.002	0.000	0.000	0.000	0.008

Notes: This table reports regressions of market-level characteristics on quantity discounts. The dependent variable in both panels is the estimated quantity discount parameter γ_v at the HTS-6 code level. Panel A uses the OLS estimate of γ_v, while Panel B uses the IV estimate. Each column regresses γ_v on a different market-level characteristic. Heteroskedastic standard errors are in parentheses. See the text for full details and sources.

that per-transaction fixed costs (e.g., for logistics, invoicing, or quality verification) are a more likely driver than initial search costs. As further evidence, Column (9) shows that markets with greater price dispersion—a common proxy for higher search frictions (Marshall, 2020)—exhibit significantly larger quantity discounts, consistent with a model where transaction costs are a key determinant of the pricing schedule.

5.4 Shipping and Transportation Costs

A key advantage of the import data is that it separately reports shipping and freight charges, allowing us to directly test for scale economies in a major component of transaction costs. If the overall quantity discounts we observe are driven by cost-side factors, we should see similar patterns in these observable cost components.

Table 7 confirms this prediction. We find substantial scale economies in shipping: regressing the log of per-unit freight costs on the log of transaction quantity, within a buyer-seller-variety relationship, yields a scale elasticity of -0.383 (Column 5). This means a 10% increase in shipment size is associated with a 3.8% decrease in per-unit shipping costs. While declared shipping charges are only 3-5% of overall transaction values, this is qualitatively consistent evidence of physical scale economies. It supports an interpretation that the overall quantity discount schedule is driven by real cost factors rather than markup adjustments.[20] This aligns with evidence that the technology of international transport is

[20]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).

characterized by significant increasing returns to scale (Hummels, Lugovskyy, and Skiba, 2009; Ganapati et al., 2024), providing a physical underpinning for the cost gradients we observe.

This elasticity is robust across specifications. Columns (4) and (5) control for the mode of transport (e.g., air, containerized sea, bulk carrier, land), isolating scale economies from differences in shipping technology. Our preferred specification in Column (5), which includes relationship, country-variety, and mode fixed effects, confirms a strong and significant elasticity of -0.383. Column (6) allows this elasticity to vary by country-variety; the mean elasticity remains large at -0.355, with a standard deviation of 0.381 across varieties (in brackets), indicating that while there is heterogeneity, shipping scale economies are a pervasive feature. The results are also robust to alternative measures of transaction size. Using shipment weight in kilograms instead of standardized quantity units yields a nearly identical elasticity of -0.378 (Column 7). Using shipment value yields a smaller but still significant elasticity (Column 8). This consistent pattern across different measures and controls provides strong, direct evidence for cost-based scale economies in a key component of transaction costs.

Table 7: Decomposition of Shipping and Other Charges

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			$\log\left(\frac{\text{Charges}}{q}\right)$				$\log\left(\frac{\text{Charges}}{\text{kg}}\right)$	$\log\left(\frac{\text{Charges}}{\text{Value}}\right)$
$\log q$	-0.508	-0.403	-0.377	-0.481	-0.383	-0.355		
	(0.000126)	(0.000116)	(0.000155)	(0.000122)	(0.000109)	[0.381]		
$\log kg$							-0.378	
							(0.000133)	
log Shipment Value								-0.168
								(0.000159)
$\log q \times$ Related Parties			-0.0577					
			(0.000232)					
R^2	0.766	0.932	0.934	0.854	0.95	0.784	0.683	0.489
Within R^2	0.219	0.185	0.188	0.215	0.196	*	0.128	0.0189
Fixed Effects								
Country-Variety	✓			✓	✓	✓	✓	✓
Relationship		✓	✓		✓			
Mode				✓	✓			

Notes: This table decomposes various measures of shipping charges. See text for details. Parentheses represent standard errors clustered at the relationship, and brackets represent the standard deviation of estimations across HS codes. See the text for full details and sources.

5.5 Discussion

The evidence strongly supports a cost-based explanation. While observable shipping charges (freight) only account for a portion of the value, they crucially reveal that the *tech-*

nological elasticity of logistics is large (≈ -0.38). Similar scale economies may apply to unobservable order-level costs, such as batch production setups, warehousing handling, and administrative processing (Arrow et al., 1951). For example, if internal batch-level costs share the same physics as external freight, they would generate the large aggregate discounts we observe. Conversely, if the remaining discount were markup-driven, we would expect it to vary with market power, which it does not.

The key empirical finding is not that we can fully explain γ_v, but that the explainable variation aligns with cost-side rather than demand-side factors. Furthermore, the direct evidence from shipping costs (Table 7), which are observable and unambiguously cost-driven, provides the strongest single piece of evidence for our interpretation.

The scale elasticity is a market-wide feature, common to all transactions, rather than a tool for price discrimination. We conclude that the firm-to-firm quantity discount elasticity is well-approximated by a market-level cost component: $\gamma_{i\to j,v} \approx \gamma^c_v$, with $\gamma^\mu_{i\to j,v} \approx 0$. Furthermore, we will conduct robustness in which we attribute all quantity discounts to markup variation.

Our finding allows us to decompose observed prices into two distinct components: a cost-driven scale effect (γ) and a residual "scale-free" price level ($\tilde{p}$). In the next section, we analyze the behavior of this residual price level to distinguish between market power and aggregate scale economies. This decomposition is crucial for understanding tariff pass-through (Section 7).

6 Aggregate Implications

Since transaction-level quantity discounts are primarily cost-driven ($\gamma \approx \gamma^c$), we can isolate a "scale-free" price, $\tilde{p}_{i\to j,v}$, representing the price level net of transaction-size effects.

6.1 The Determinants of Scale-Free Prices

Relationship Scale Economies Do transaction-level economies aggregate to the firm or relationship level? Recent policy literature often estimates scale economies at the firm or higher levels (Lashkaripour and Lugovskyy, 2023; Bartelme et al., 2025; Farrokhi and Soderbery, 2020; Duarte, Magnolfi, Quint, Sølvsten, and Sullivan, 2025). Going beyond correlations, we test whether large buyers pay lower prices because of aggregate volume (firm scale) or simply because they make larger individual shipments (transaction scale).

We employ a first-difference strategy using 2011-2012 data to identify the relationship-

level supply elasticity:

$$\Delta \log \tilde{p}_{i \to j,v} = \beta \Delta \log q_{i \to j,v} + FE_{i,v} + \epsilon_{ij}, \tag{15}$$

where $\Delta \log q_{i \to j,v}$ is the change in total quantity purchased by buyer j from seller i. We control for seller-variety fixed effects ($FE_{i,v}$) to account for supply-side shocks. We instrument for the change in quantity using a shift-share instrument based on aggregate demand shocks in downstream industries.[21]

Table 8 compares regressions using raw average prices ($\Delta \log p$) versus scale-free prices ($\Delta \log \tilde{p}$).

When using raw prices (Columns 1, 3, 5, 7), we find a significant negative relationship between price and quantity, suggesting a level of aggregated scale economies. However, when we use the scale-free price $\tilde{p}$ (Columns 2, 4, 6, 8), this relationship largely disappears or is substantially attenuated. This implies that apparent aggregate scale economies are a mechanical reflection of transaction-level discounts. Large buyers pay less primarily because they transact in larger batches.

Our estimates align with Lashkaripour and Lugovskyy (2023) who find aggregate scale of $\gamma \approx 0.2$ in Colombia, and with Bartelme et al. (2025) who find scale of $\gamma \approx 0.2$ globally. Macroeconomic models assuming firm-level scale economies may mis-specify the mechanism if the true driver is at the transaction-level. We replicate these findings with domestic data in Appendix C.2.1. In Appendix B.5, we also decompose scale-free prices in the cross section.

Validating the Decomposition If our decomposition is correct, "level" effects like market power may reside in $\tilde{p}$, not γ. Table 9 confirms this. This is consistent with the literature on how there is wide variation in both how exporters react to various shocks (Berman, Martin, and Mayer, 2012) and in what determines price variation across relationships (Manova and Zhang, 2012; Kugler and Verhoogen, 2012; Basu and Fernald, 1997).

Panel A shows correlations between prices and measure of market power within Variety-Country. Panel B shows the same correlations with more demanding fixed effects. Depending on the specification, there are seller-variety and/or buyer-variety-country fixed effects.

Unlike the discount elasticity, the scale-free price $\tilde{p}$ is significantly correlated with bilateral market shares and seller concentration. Market power shifts the *intercept* of the

[21] Given what we have disclosed from Census we use 2011-2012 data here instead of 2016-2017.

Table 8: Relationship Scale Economies - Bilateral Price Changes

	(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-Variety	✓	✓						
Seller-Variety			✓	✓	✓	✓	✓	✓
First Stage F							1000	
Instruments							✓	✓
IV Sample					✓	✓	✓	✓

Notes: This table reports regressions of the change in aggregate quantities on the change in price. Odd columns use aggregate prices ($\Delta \log p$). Even columns use the scale-free price ($\Delta \log \tilde{p}$), which adjusts for transaction-level quantity discounts. The last two columns instrument for the change in quantity using the shift-share instrument described in the text. See the text for full details and sources.

pricing schedule but does not alter its *slope*. Large firms may negotiate lower base prices, but the marginal incentive regarding shipment size is determined by common technological factors.

6.2 From Micro-Scale to Macro-Pass-Through

Our analysis reveals a stark divergence between reduced-form and structural interpretations of price data. At the micro-level, we find that "scale economies" are largely a phenomenon of shipment size rather than firm size. At the macro-level, this distinction is critical for understanding the incidence of trade shocks.

If scale economies were driven by firm-level market power or aggregate production functions, a tariff shock might be absorbed by a reduction in firm-level markups or a movement along a firm-level cost curve. However, these scale economies can be driven by transaction-level logistics (e.g., fixed ordering costs), not firm-level production functions. As tariffs raise the holding cost of inventory, firms optimally reduce their order sizes. This reduction in transaction size forces buyers up the quantity discount schedule. Consequently, observed unit prices rise for two distinct reasons: the direct cost of the tariff, and the mechanical loss of scale economies. Standard pass-through estimates conflate these two effects, mistakenly attributing the mechanical price increase to strategic pass-through. In the next section, we show that once we correct for this mechanical

Table 9: Scale-Free Price Level Variation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable					$\log(\tilde{p})$				
Independent Variable	Related Party	$\log(q)$ Relationship	$\log(q)$ Seller	$\log(q)$ Buyer	Output Share	Input Share	log(Rel) Length	log(Sellers)	log(Buyers)
Panel A: Baseline Level Effects									
Coefficient	0.0555	-0.172	-0.137	-0.0877	-0.154	-0.125	0.000909	0.0175	0.033
	(0.00437)	(0.00292)	(0.00238)	(0.00178)	(0.00555)	(0.00538)	(0.00153)	(0.00270)	(0.00411)
R^2	0.0863	0.217	0.177	0.133	0.0887	0.0878	0.086	0.0863	0.0864
Within R^2	0.000425	0.144	0.0995	0.0514	0.003	0.00206	0.000000706	0.000429	0.000546
Fixed Effects									
Product-Country	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel B: Complex Fixed Effects									
Coefficient	-0.011	-0.0289	-0.0501	-0.0281	-0.131	-0.0454	-0.000676	-0.0169	0.0162
	(0.00250)	(0.00152)	(0.000860)	(0.000724)	(0.00646)	(0.00516)	(0.00254)	(0.00140)	(0.00140)
R^2	0.854	0.855	0.686	0.749	0.854	0.854	0.854	0.747	0.68
Within R^2	0.000123	0.009	0.0189	0.00732	0.00342	0.000866	0.000000189	0.000655	0.0002
Fixed Effects									
Seller-Variety	✓	✓			✓	✓		✓	
Buyer-Variety-Country	✓	✓	✓	✓	✓	✓	✓		✓

Notes: This table reports regressions of the scale-free price level on various relationship and market characteristics. Panel A includes product-country fixed effects. Panel B includes more complex fixed effects, including seller-variety and buyer-variety-country fixed effects. Clustered source-variety standard errors are in parentheses. See the text for full details and sources.

"scale effect," the true pass-through of the 2018-2019 US tariffs is significantly lower than previously thought.

7 Tariff Pass-Through

Cost-driven ($\gamma \approx \gamma^c$) quantity discounts fundamentally alter the interpretation of aggregate price movements. In standard models, incomplete pass-through implies markup compression. Our results suggest a countervailing force: if tariffs reduce trade volumes, they push firms up their average cost curves, mechanically raising unit prices.

This implies that standard estimates of unit-value pass-through do not alone measure incidence. A tariff-induced price increase could reflect two distinct phenomena: the tax itself, or the loss of scale economies due to shrinking transaction sizes. By failing to account for the latter, we risk underestimating if foreign firms are actually cutting markups (or scale-free prices) to maintain market share, hiding a substantial welfare loss borne by foreign exporters.

Table 10: Summary of US Import Data

Origin	Year	Value (FOB)	Duty	Shipping Charges	Tariff Rate	Shipping Rate
		(Billion USD)			(Applied %)	(Effective %)
All	2017	1,796	29.1	52.0	1.6%	2.9%
All	2018	2,005	42.9	57.2	2.1%	2.9%
All	2019	2,274	64.8	63.2	2.9%	2.8%
All	2020	2,316	56.5	67.1	2.4%	2.9%
China	2017	363	11.6	14.3	3.2%	3.9%
China	2018	398	19.6	16.5	4.9%	4.1%
China	2019	425	40.2	17.4	9.5%	4.1%
China	2020	460	37.0	20.5	8.0%	4.5%
Other	2017	1,434	17.5	37.7	1.2%	2.6%
Other	2018	1,606	23.3	40.7	1.5%	2.5%
Other	2019	1,848	24.6	45.8	1.3%	2.5%
Other	2020	1,855	19.5	46.6	1.1%	2.5%

Notes: Data are from the US Census LFTTD and cover US imports exceeding the de minimis threshold. Value (FOB) is the free-on-board value of imports. Duty is the total customs duties paid. Shipping Charges include freight and insurance costs. The Tariff Rate is the applied rate, calculated as Duty divided by Value. The Shipping Rate is calculated as Shipping Charges divided by Value. See the text for full details and sources.

7.1 2018–2019 US Tariffs

The 2018–2019 trade war provides a quasi-experiment for studying tariff pass-through. During this period, the United States imposed substantial tariffs on a wide range of imported goods, with a primary focus on products originating from China. The aggregate impact of these tariffs is evident in Table 10. Total duties collected on US imports more than doubled from $29.1 billion in 2017 to $64.8 billion in 2019, raising the average applied tariff rate from 1.6% to 2.9%. The burden fell disproportionately on imports from China, where the average tariff rate nearly tripled from 3.2% to 9.5% over the same period. In contrast, the average tariff rate on imports from the rest of the world remained stable at around 1.3%.

The standard approach to measuring tariff pass-through treats observed unit price as the relevant object for welfare analysis. However, when transaction-level quantity discounts are pervasive, this approach conflates two distinct effects: changes in the scale-free price schedule $\tilde{p}$ and compositional shifts in transaction sizes. Our framework from Section 2.2 clarifies this distinction. The total change in observed prices reflects both the shift in the underlying cost and markup structure (captured by $d\tilde{p}/d(1+\tau)$) and the adjustment in transaction quantities (captured by $\gamma \cdot dq/d(1+\tau)$). Moreover, to measure aggregate pass-through, we must aggregate transaction-level prices to construct unit values, accounting for both the intensive margin (changes in the size of existing transactions) and

the extensive margin (entry and exit of trading relationships). Only by separating these components can we accurately measure the division of the tariff burden between buyers and sellers.

This distinction matters fundamentally for incidence. As we showed in our conceptual framework, the change in buyer surplus depends on shifts in the scale-free price $\tilde{p}$ weighted by quantities purchased (Equation (2)), while changes in producer surplus require decomposing both revenue and cost effects (Equation (3)). When transaction sizes decline in response to tariffs—as our data show they do—standard measures of pass-through will overstate the extent to which sellers can shift the tariff burden to buyers. This is because smaller transactions occur at higher unit prices due to scale economies, creating the appearance of greater pass-through even when the underlying price schedule has shifted down substantially.

Our empirical strategy leverages the transaction-level variation we document to construct scale-free price measures (an implicit unit-value price index). By recovering the scale elasticity γ from within-relationship variation (Section 4.2) and using it to compute scale-free prices $\tilde{p}$, we isolate the true shift in the price schedule from compositional effects. This allows us to decompose the incidence of the 2018 tariffs into three components: the mechanical effect of the tariff itself, the adjustment in seller markups and costs (captured by changes in $\tilde{p}$), and the quantity response both at the transaction level and through entry and exit. Our findings reveal that accounting for transaction-level scale economies reduces estimated pass-through from near unity to approximately 60 percent, implying that foreign exporters absorbed a substantially larger share of the tariff burden than conventional estimates suggest.[22]

Essentially, we empirically differentiate reduced-form versus structural estimates of pass-through (MacKay, Miller, Remer, and Sheu, 2014). We do find that at both the aggregate and transaction levels, we can match the reduced-form findings of the literature (Fajgelbaum et al., 2020). Considering an aggregate variety, US tariffs on imports exhibit close to full reduced-form pass-through of tariffs to unit wholesale prices. 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.

Empirical Framework for Tariff Pass-Through with Transaction-Level Scale Economies. At the transaction level, we start off with the tariff-inclusive transaction price within a re-

[22]Unlike our main regressions where prices were demeaned, here we keep them at their initial levels.

lationship:

$$\log p_{i\to j,t,v} = \underbrace{\log\left(1+\tau_{t,v}\right) + \log \mu_{i\to j,t,v} + \log \tilde{c}_{i\to j,t,v}}_{\log \tilde{p}_{t,v}} + \gamma \log q. \tag{16}$$

Prices are a combination of tariffs (ad-valorem), scale-free markups and marginal costs, and a final scale effect. From the analysis in Section 3, we show that this scale effect is driven by underlying costs and we will consider them invariant to tariff levels.[23]

For any given transaction t from seller i to buyer j at date d, as shown in section 4.2, there is a scale pricing elasticity γ. When considering tariffs, tariffs have log-separable effects on duty-inclusive prices:

$$\frac{\partial \log p_{i\to j,t,v}\left(q\right)}{\partial \log\left(1+\tau_{t,v}\right)} = \underbrace{1 + \frac{\partial\left[\log \mu_{i\to j,t,v} + \log \tilde{c}_{i\to j,t,v}\right]}{\partial \log\left(1+\tau_{t,v}\right)}}_{\tilde{\rho}=\partial \log \tilde{p}/\partial \log(1+\tau)} + \gamma \frac{\partial \log q}{\partial \log\left(1+\tau_{t,v}\right)}. \tag{17}$$

The first element on the right-hand side is the effect of a tariff on a scale-independent price. This can be further decomposed into a markup effect μ, and a scale-free cost effect $\tilde{c}$. The second element is the effect of a tariff on the scale of an individual transaction.[24]

7.1.1 Within-Relationship Transaction-Level Pass-Through

Looking at data on tariffs imposed by the US on its own imports, Table 11 shows the results of considering Equation (17). We consider prices, not only within a buyer-seller-variety relationship, but relative to variety-time trends (to partially account for the supply shock effect in Gertler (2023) and the diversions seen in Fajgelbaum, Goldberg, Kennedy, Khandelwal, and Taglioni (2024) and Farrokhi and Soderbery (2020).

Column (1) shows that the pass-through of tariffs to observed transaction prices in 2018 is nearly complete, with a coefficient of 0.956. This finding, which aligns with the reduced-form results in the literature, suggests that for every 10% increase in the tariff rate, the post-tax price paid by the buyer increases by almost 10%. However, this masks a significant quantity response. Column (5) reveals that a 10% tariff rate increase leads to a 6% decrease in the size of individual transactions.[25]

[23] Table 15 shows that γ is invariant to tariffs.

[24] Decomposing incomplete pass-through—distinguishing between costs, markup adjustment, and mechanical invoicing effects—is a focus of the exchange rate literature (Goldberg and Hellerstein, 2008).

[25] In Appendix B.4, we replicate this analysis using the tariffs imposed by the US on consumer goods in the last quarter of 2019, finding similar results. However, the estimates are noisier due to the shorter time frame before the onset of the COVID-19 pandemic.

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

	(1) $\log(p)$	(2) $\log(\tilde{p}^{OLS})$	(3) $\log(\tilde{p}^{IV})$	(4) $\log(pq)$	(5) $\log(q)$	(6) $\log(p)$
log(1+ Tariffs Applied)	0.956	0.780	0.800	0.357	-0.600	
	(0.0424)	(0.0282)	(0.0360)	(0.0438)	(0.0603)	
log(1+ Tariffs Statutory)						0.920
						(0.0417)
R^2	0.965	0.967	0.996	0.725	0.912	0.965
Within R^2	0.000746	0.000809	0.000358	0.0000486	0.000103	0.00045
Fixed Effects	Buyer-Seller-Variety, Variety-Year-Month					

Notes: This table reports transaction-level duty-inclusive tariff pass-through estimates for continuing buyer-seller-variety relationships in 2018. All specifications include buyer-seller-variety and variety-year-month fixed effects. Standard errors are clustered at the relationship level. Column (1) shows the pass-through of applied tariffs (duty paid) to observed transaction prices. Columns (2) and (3) report pass-through to scale-free prices ($\tilde{p}$), constructed using the OLS and IV estimates of the scale elasticity γ from Section 4.2, respectively. These columns isolate the change in the price schedule from compositional effects. Column (4) reports the effect on transaction value, while Column (5) shows the response of transaction quantity. Column (6) shows the pass-through of statutory tariffs. See the text for full details and sources.

Our framework allows us to disentangle these effects. Columns (2) and (3) report the pass-through to the scale-free price, $\tilde{p}$, which isolates the shift in the underlying price schedule. Using our preferred IV estimate for the scale elasticity, we find that the pass-through to the scale-free price is only 0.80 (Column 3). This implies that the composite of seller markups and costs fell by 20% in response to the tariff. If we assume that marginal costs for a standardized order size did not decrease, this suggests that foreign exporters absorbed a significant portion of the tariff burden by reducing their markups. Had transaction sizes not fallen, the observed pass-through would have been only 80%, below the near-unity estimates that ignore quantity adjustments. This highlights a key difference from frameworks like Fajgelbaum et al. (2020), which, by assuming perfect competition and no quantity discounts ($\gamma = 0$), would interpret the full price increase as being borne by US buyers.

The decomposition in Equation (17) provides a clear interpretation of these results. The coefficient in Column (1) represents the total pass-through to the observed price, $\partial \log p / \partial \log(1+\tau)$. The coefficient in Column (3) is the pass-through to the scale-free price, $\partial \log \tilde{p} / \partial \log(1+\tau)$. The difference between these two, $0.956 - 0.800 = 0.156$, represents the quantity adjustment effect, $\gamma \cdot \partial \log q / \partial \log(1+\tau)$. Using our estimated γ of -0.29 and the quantity response from Column (5) of -0.604, this effect is $(-0.29) \times (-0.600) = 0.174$, which closely matches the observed difference. Furthermore, the pass-

through to the scale-free price can be decomposed into the mechanical tariff effect (1) and the change in the composite of seller markups and costs. The estimate of 0.8 implies that this composite fell by 20 percent ($0.8 - 1 = -0.2$), indicating that foreign exporters absorbed a substantial portion of the tariff by reducing their pre-tax prices.

However, these results are based on continuing relationships and thus only capture the intensive margin of adjustment among stable relationships. Essentially, this is pass-through conditional on survival and constant order frequency and not an estimate of aggregate pass-through. To assess the full welfare incidence of the tariffs, we must aggregate these transaction-level findings to account for compositional changes.

7.1.2 Aggregate Pass-through

Moving beyond transaction-level pass-through, we now examine how quantity discounts affect aggregate unit values—the standard object used in prior tariff incidence studies. Aggregation introduces two key margins of adjustment: the intensive margin (changes in transactions and transaction sizes within and across continuing relationships) and the extensive margin (entry and exit of trading relationships). To properly measure aggregate pass-through, we must account for both how the scale-free price $\tilde{p}$ shifts in response to tariffs and how the composition of transactions changes. We present two complementary approaches: first, a scale-free aggregation that uses our estimated scale elasticity γ to construct standardized unit values from transaction-level data; and second, a first-order approximation using only aggregate data that yields quantitatively similar results and can be implemented without access to microdata.

We aggregate these scale-free prices across all transactions for a given product-origin-date combination, weighting by transaction quantities to obtain an aggregate scale-free unit value $\tilde{p}_{o,d,v}$, from origin country o to the destination d (the US in our case) for variety v. This approach allows us to isolate the shift in the underlying price schedule from compositional changes in transaction sizes and counts.

To construct scale-free unit values, we follow two steps. We first adjust transaction-level prices to remove the effect of quantity discounts using our estimated scale elasticity γ_v. This yields a scale-free price $\tilde{p}_{i \to j,t,v}$ for each transaction that reflects what the price would be if the transaction size were one unit:

$$\tilde{p}_{i \to j,t,v} = \frac{p_{i \to j,t,v}}{q_{i \to j,t,v}^{\gamma_v}}, \tag{18}$$

where γ is recovered through the mechanism in section 4.2.[26]

We create a duty-inclusive unit value in the spirit of aggregate regressions in Fajgelbaum et al. (2020); Amiti et al. (2019):

$$p_{o,d,v} = \sum_{i \in o, t \in d} p_{i \to j,t,v} q_{i \to j,t,v} / \sum_{i \in o, t \in d} q_{i \to j,t,v}. \tag{19}$$

Our scale-free prices aggregates across all buyers and sellers, where shipment of variety v originates in country o at date d, aggregated at the monthly level:

$$\tilde{p}_{o,d,v} = \sum_{i \in o, t \in d} \tilde{p}_{i \to j,t,v} q_{i \to j,t,v} / \sum_{i \in o, t \in d} q_{i \to j,t,v}. \tag{20}$$

With this approach, we can decompose the duty-inclusive aggregate price for a product as:

$$\log p_{o,d,v} = \log \tilde{p}_{o,d,v} + \log \tilde{q}_{o,d,v}, \tag{21}$$

where the residual $\tilde{q}_{o,d,v}$ captures the aggregate quantity composition effect. This term reflects how the distribution of transaction sizes affects the observed unit value, even holding the underlying price schedule $\tilde{p}$ fixed. The scale-free price $\tilde{p}_{o,d,v}$ can be interpreted as the tariff-inclusive composite of scale-free markups and marginal costs: $\tilde{p}_{o,d,v} = (1 + \tau_{o,d,v})(\mu\tilde{c}_{o,d,v})$, where $\mu\tilde{c}_{o,d,v}$ represents the quantity-weighted average of seller markups and costs, net of tariff charges.

Table 12 regresses each of these components, both the left- and right-hand sides, on changes in applied tariffs. We consider different levels of aggregation, the baseline in Panel (A) considers just monthly 2018 data. We later consider robustness in Panels (B)-(E), aggregating up to the yearly level.

Column (1) replicates the baseline findings, pass-through, not just at the transaction level, but the product-origin level is effectively unity. Columns (2) and (3) adjust for quantity discounts using our OLS and IV estimates for the scale elasticity γ_v within variety.

We can further use the difference between the aggregate un-adjusted regression in column (1) with the scale-free results in columns (2)-(5) to consider the relative effects on sellers. In our baseline comparing to our IV results, a 10% increase in tariffs decreases the prices received by sellers by 4.0%. Without the scale adjustment, the prices would have increased by 0.5%.

The difference between the unweighted transaction-level estimate (0.80) and the volume-weighted aggregate estimate (0.60) indicates that compositional effects are significant.

[26] Alternatively, we can use $\gamma_{i \to j,v}$.

Larger transactions, which benefit more from quantity discounts, tend to shrink more in response to tariffs. This compositional shift amplifies the mechanical price increase due to reduced transaction sizes, leading to a lower aggregate pass-through rate.

The microdata founded scale elasticities play a crucial role. But how important is microdata for this exercise? As an alternative, we bound our results using a much more straightforward decomposition, conditional on recovering scale elasticities from the microdata.

As an alternative to Equation (21), we can interpret results using a much more straightforward decomposition,

$$\log p_{o,d,v} = \log pq_{o,d,v} - \log q/T_{o,d,v} - \log T_{o,d,v}, \tag{22}$$

where T denotes the number of transactions. Here, we decompose the aggregate unit value into three components: the total payments received by sellers for variety v from origin o at date d (i.e., total revenue) denoted by $pq_{o,d,v}$, the average order size captured by $q/T_{o,d,v}$, and the number of transactions $T_{o,d,v}$.

Table 12 in columns (5-8) regresses tariff rates on each of these components. Column (5) indicates a 10% increase in tariffs leads to an 8.7% decrease in payments. Column (7) shows that a 10% increase in tariffs leads to a 15.5% decrease in average order size. Column (8) shows that a 10% increase in tariffs leads to a 3.6% decrease in the number of transactions.

In particular the second term, $\log q/T_{o,d,v}$, is informative of the level of pass-through netting out quantity changes, in combination with our γ_v from Section 4.2. As an approximation, if we assume the change in average order size is reflective of the weighted average change in order size at the transaction level, we can use our earlier estimate of γ to back out the change in scale-free prices:

$$\frac{d\log \tilde{p}_{o,d,v}}{d\log(1+\tau_{o,d,v})} \approx \frac{d\log p_{o,d,v}}{d\log(1+\tau_{o,d,v})} - \gamma\frac{d\log\left(q/T\right)_{o,d,v}}{d\log(1+\tau_{o,d,v})} = 1.05 - (-1.55)*(-0.29) = 0.60.$$

Essentially, we recover columns (2)-(4) to a first order approximation. For a consistent order size, tariffs are only passed through at the aggregate level at 60%.[27] As with before, we can decompose this into a tariff direct effect (mechanically 1) and a markup and cost effect. If we assume that costs did not fall during the tariff war, then aggregate markups fell 40%.

[27]Due to Jensen's inequality, the logarithm of a sum is not the sum of a logarithm (as log is a concave function), thus requiring the use of detailed transaction-level data.

The discrepancy between the relationship-level pass-through (Table 11) and the aggregate results stems from the extensive margin. As shown in Column (7), the number of transactions declines significantly in response to tariffs. This reduction in order frequency, combined with the reduction in order size, compounds the welfare loss. While continuing relationships see a pass-through of roughly 0.80 to scale-free prices, the exit of relationships and the consolidation of orders implies that the aggregate average unit value rises by more than the intensive-margin estimate would suggest. This aligns with work emphasizing the extensive margin of global sourcing, where fixed costs of importing lead to sensitivity in the measure of active suppliers (Antras, Fort, and Tintelnot, 2017).

Correcting for Quality For robustness, we also consider a Feenstra (1994)-style CES price index within each origin-destination-variety combination to account for the extensive margin of entry and exit. We use data on the elasticity of substitution from a structural model of supply and demand.

We construct an exact CES price index to account for substitution bias and variety turnover (entry and exit). Using the Sato-Vartia index for continuing goods and the Feenstra adjustment for the extensive margin, we find that standard unit-value indices may overstate inflation by failing to account for substitution towards cheaper varieties. However, for the medium-run pass-through estimates in Table 12 column (4), the divergence between the CES index and our simple average is minimal, suggesting that substitution bias is not the primary driver of our results at this horizon.

Event Horizons In Table 12 Panels (B)-(E), we consider different time periods and aggregation horizons. We first include all data from 2017-2019 at the monthly level (B), before aggregating to the quarterly (C) and year levels (D). Lastly we take the yearly data from 2017 and do a two-year difference to 2019 (E).

Echoing the robustness of the unadjusted pass-through in Fajgelbaum et al. (2020), our baseline results in column (2)-(4) adjusting our unit price indices for quantity discounts are largely consistent across panels, with measured pass-through between 0.4 and 0.7, with OLS scale elasticities at the higher end and the CES-adjustment at the lower end.

Decomposing Aggregate Pass-through In aggregate, pass-through is the combination of multiple forces, the pass-through of continuing firms, and the net effect of entrants and exiting relationships.

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

Table 12: Decomposing Aggregate Tariff Pass-Through by Methodology and Time Period

	(1) $\log(P)$	(2) $\log(\tilde{p}_{\gamma\text{OLS}})$	(3) $\log(\tilde{p}_{\gamma\text{IV}})$	(4) $\log(\tilde{p}_{\text{CES}})$	(5) log(Value)	(6) $\log(Q)$	(7) $\log(Q/T)$	(8) $\log(T)$
Panel A: Monthly 2018 Data								
$\log(1+\text{Tariffs})$	1.047	0.804	0.604	0.573	-0.867	-1.914	-1.551	-0.362
	(0.0546)	(0.0414)	(0.0447)	(0.0418)	(0.0709)	(0.0897)	(0.0755)	(0.0394)
r^2	0.919	0.942	0.951	0.951	0.876	0.888	0.913	0.918
Within R^2	0.000478	0.000312	0.000223	0.000223	0.000155	0.000301	0.000757	0.000155
Panel B: Monthly 2017-2019 Data								
$\log(1+\text{Tariffs})$	1.038	0.547	0.537	0.504	-1.479	-2.517	-1.697	-0.820
	(0.0302)	(0.0226)	(0.0238)	(0.0466)	(0.0415)	(0.0526)	(0.0424)	(0.0258)
r^2	0.897	0.927	0.938	0.686	0.861	0.891	0.88	0.897
Within R^2	0.00065	0.000357	0.000278	0.000276	0.00125	0.00182	0.00111	0.00112
Panel C: Quarterly 2017-2019 Data								
$\log(1+\text{Tariffs})$	1.069	0.584	0.568	0.436	-1.669	-2.739	-1.751	-0.988
	(0.0362)	(0.0276)	(0.0299)	(0.0260)	(0.0507)	(0.0639)	(0.0503)	(0.0318)
r^2	0.909	0.934	0.945	0.65	0.892	0.909	0.896	0.918
Within R^2	0.000672	0.000386	0.000295	0.000456	0.00151	0.00205	0.00118	0.0014
Panel D: Yearly 2017-2019 Data								
$\log(1+\text{Tariffs})$	1.178	0.68	0.645	0.481	-2.032	-3.21	-1.886	-1.324
	(0.0578)	(0.0469)	(0.0522)	(0.0240)	(0.0830)	(0.104)	(0.0809)	(0.0528)
r^2	0.937	0.954	0.962	0.578	0.934	0.94	0.928	0.949
Within R^2	0.000808	0.000505	0.000367	0.000801	0.0021	0.00267	0.00138	0.00212
Panel E: Two Year Annual Change 2017 to 2019 Data								
$\log(1+\text{Tariffs})$	1.093	0.645	0.602	0.533	-2.189	-3.283	-1.737	-1.546
	(0.0654)	(0.0516)	(0.0591)	(0.0303)	(0.0938)	(0.118)	(0.0891)	(0.0630)
r^2	0.949	0.963	0.969	0.544	0.938	0.947	0.939	0.95
Within R^2	0.00105	0.000684	0.000471	0.00122	0.00321	0.00388	0.00172	0.00359
Fixed Effects	Variety-Country Origin, Time-Month							

Notes: This table reports aggregate tariff pass-through estimates at the origin-destination-variety level using different methodologies and time periods. Each panel represents a different time aggregation: Panel A uses monthly data from 2018, Panel B uses monthly data from 2017-2019, Panel C uses quarterly data from 2017-2019, Panel D uses yearly data from 2017-2019, and Panel E uses two-year changes from 2017 to 2019. Column (1) shows the pass-through of applied tariffs to observed unit values. Columns (2) and (3) report pass-through to scale-free unit values, constructed using the OLS and IV estimates of the scale elasticity γ_v, respectively. Column (4) presents pass-through estimates using a CES price index to account for quality adjustments. Columns (5) to (8) decompose the aggregate unit value into total value, total quantity, average order size, and number of transactions. All specifications include variety-country origin and time-month fixed effects. Standard errors are clustered at the variety-country origin level. See the text for full details and sources.

medium run difference between 2017 and 2019.[28]

In Table 13, we run effectively the same decomposition as in the previous subsection, but in first differences, as there are only two periods. In Panel A, we consider continuing

[28]This ignores the set of consumer products tariffed in late 2019 on the eve of the COVID-19 pandemic.

trading partners that account for approximately 2/3 of trading volume. In this case, aggregate pass-through is only 0.87 (Col 1), with a large order size and order number effect (Col 5-6). However, our scale-corrected pass-through estimates are .60 and .57 (Col 2-3).

Panel B considers the net effect of entrant and exiting trading pairs. Aggregate pass-through, under the strict assumption of comparability, implies that pass-through is more than full (1.2) for this set of firms (Col 1) with even larger large order size and order number effects (Col 5-6). But these differences with continuing relationship are minimized when considering the scale-corrected pass-through estimates of .64 and .60 (Col 2-3).[29]

There are massive compositional changes induced by tariffs, but estimates correcting for quantities aligned our estimates of pass-through between various samples.

Unification With Exchange Rate Pass-Through The nearly complete pass-through of tariffs into aggregate unit import prices (Amiti et al., 2019; Fajgelbaum et al., 2020; Cavallo et al., 2021) stands in stark contrast to the Exchange Rate Pass-Through (ERPT) literature, which typically finds much lower pass-through rates into import prices, often in the range of 0.4 to 0.6, even in the medium run (Gopinath et al., 2010). This discrepancy presents a puzzle: why would foreign exporters absorb exchange rate shocks but fully pass on tariff shocks?

A key distinction between these literatures lies in the data used. Tariff studies often rely on customs unit values, which are calculated as total value divided by total quantity. As we have shown, unit values mechanically incorporate changes in order size and the resulting quantity discounts. In contrast, the ERPT literature, such as Gopinath et al. (2010), typically utilizes BLS import price indices. These indices are constructed from survey data that track the price of specific items over time, and in principle condition on the quantity discount channel we highlight.[30]

When we correct for scale effects and isolate the "scale-corrected" pass-through, analogous to controlling for item specifics and quantity, we estimate a pass-through rate of approximately 0.60 - much closer to the standard ERPT estimates. The apparent difference between tariff and exchange rate pass-through may be driven by the measurement of prices: unit values capture the endogenous response of quantities and discounts, whereas price indices isolate the pure price change.

This lines up with literature that emphasizes strategic complementarities and variable markups as the primary drivers of incomplete pass-through (Atkeson and Burstein,

[29]These estimates use the same scale elasticities as in Panel A, which are derived from our within-relationship IV estimates for γ_v.

[30]Although note that for their import price index, the BLS recently switched to using administrative trade data for 40 percent of their sample (U.S. Bureau of Labor Statistics, 2025).

2008; Amiti et al., 2014). Both tariff and exchange rate shocks affect the underlying cost structure of exporters, leading to adjustments in markups. When controlling for quantity effects, both types of shocks exhibit similar pass-through patterns, suggesting that the underlying market dynamics are consistent across these different types of trade shocks.

Table 13: Decomposing Medium Run Tariff Pass-Through

Panel A: Continuing Relationships from 2017 to 2019

	(1) $\log(P)$	(2) $\log(\tilde{p}_{\gamma\text{OLS}})$	(3) $\log(\tilde{p}_{\gamma\text{IV}})$	(4) $\Delta\log(\text{Value})$	(5) $\Delta\log(Q)$	(6) $\Delta\log(Q/T)$	(7) $\Delta\log(T)$
$\Delta\log(\text{Tariffs})$	0.871	0.603	0.572	-0.861	-1.731	-0.73	-1.002
	(0.0607)	(0.0450)	(0.0517)	(0.0863)	(0.104)	(0.0763)	(0.0672)
R^2	0.129	0.123	0.126	0.127	0.133	0.125	0.136
Within R^2	0.00189	0.00165	0.00113	0.000916	0.00254	0.000843	0.00204

Panel B: Other Firms

	(1) $\log(P)$	(2) $\log(\tilde{p}_{\gamma\text{OLS}})$	(3) $\log(\tilde{p}_{\gamma\text{IV}})$	(4) $\Delta\log(\text{Value})$	(5) $\Delta\log(Q)$	(6) $\Delta\log(Q/T)$	(7) $\Delta\log(T)$
$\Delta\log(\text{Tariffs})$	1.222	0.637	0.6	-2.857	-4.079	-2.278	-1.801
	(0.0964)	(0.0697)	(0.0779)	(0.114)	(0.155)	(0.121)	(0.0757)
R^2	0.0887	0.0896	0.0846	0.119	0.112	0.0949	0.134
Within R^2	0.00108	0.000559	0.000397	0.00421	0.00464	0.00238	0.00378
Fixed Effects				Variety-Year			

Notes: This table decomposes medium-run tariff pass-through by relationship status. Panel A reports estimates for continuing relationships. Panel B reports estimates for the net effect of entering and exiting relationships. All specifications are in first differences between June 2018 and June 2019 and include variety-year fixed effects. Standard errors are clustered at the variety-origin level. See the text for full details and sources.

7.2 Revisiting Incidence

Overall, using a conventional unit price index—as opposed to one adjusted for transaction quantities—would suggest near-complete pass-through. However, our scale-free results indicate a net pass-through of approximately 50-60%, with the difference driven by substantial decreases in average order size and the number of orders.

Having established that pass-through is incomplete and quantity responses are large, we now combine these estimates to assess the overall welfare burden. We return to the framework from Section 2.2. While a full welfare analysis typically requires a structural supply and demand system, Weyl and Fabinger (2013) show that marginal effects can be recovered using sufficient statistics. We extend this logic to our setting, incorporating the imperfect competition wedges we have empirically identified.

We model seller prices p_s, revenues R, and costs C as:

$$p_s = \frac{\tilde{p}}{1+\tau} q^{\gamma_c}, \quad R = \frac{\tilde{p}}{1+\tau} q^{\gamma_c+1}, \quad C = \tilde{c} q^{\gamma_c+1}$$

Locally, the changes in revenue and cost with respect to tariffs are:

$$\frac{dR(q)}{d(1+\tau)} = \left[\frac{d\log\tilde{p}}{d\log(1+\tau)} + (\gamma_c+1) \frac{d\log q}{d\log(1+\tau)} - 1 \right] \frac{R}{1+\tau}$$

$$\frac{dC(q)}{d(1+\tau)} = \left[\frac{d\log\tilde{c}}{d\log(1+\tau)} + (\gamma_c+1) \frac{d\log q}{d\log(1+\tau)} \right] \frac{R}{\tilde{\mu}(1+\tau)}$$

Note that terms involving γ^μ drop out due to our finding that quantity discounts are cost-driven ($\gamma^\mu \approx 0$). To complete this calculation, we need two additional pieces of information: the elasticity of scale-free costs $\tilde{c}$ to tariffs, and the baseline level of markups $\tilde{\mu}$. In addition, we need to test the implicit assumption that γ is invariant to tariffs.

Elasticity of Scale-Free Cost to Tariffs Understanding how scale-free costs change due to the trade war is not trivial. In principle, this needs to be known for every existing relationship. As we are primarily interested in incidence, we proceed under the assumption that the level of marginal costs is invariant to the destination, i.e., $\tilde{c}_{odv} = \tilde{c}_{ov}$, and that markups to foreign markets are unaffected.

We assume that products sold to the United States have the same cost structure as those sold to the rest of the world and that markups to 3rd parties are also unchanged. Essentially, the idea is that goods sold to Europe and America use the same mix of inputs. We test this by examining unit prices of exports to third-party countries:

$$\frac{d\log\tilde{p}_{o,-US,v}}{d\log(1+\tau_{o,US,v})} = \frac{d\log\tilde{c}_{o,US,v}}{d(1+\tau_{o,US,v})},$$

where we look at shipments from origin o to non-US destinations $-US$ for variety v. As we do not have transaction-level data, we leverage results from the prior section and consider using aggregate shipment quantities to control for scale-effects.

Table 14 highlights the recovery of the pass-through of US tariffs on goods exported via third party trade, controlling for the transaction-level aggregation in Section 6. We consider data from 2017 (pre-tariff) and 2019 (post-tariff). Columns (1) and (2) conduct a first difference regression with varying levels of fixed effects and controlling for aggregate scale economies. We find that US tariffs did not change $\tilde{p}$ for third countries. Column (3) conducts a two-stage regression, first netting out scale economies and then analyzing the

Table 14: Decomposing US Tariff Pass-Through on Foreign Prices

	(1) $\Delta \log \tilde{p}_{i,j\neq US,v}$	(2) $\Delta \log \tilde{p}_{i,j\neq US,v}$	(3) $\Delta \log \tilde{p}_{i,j\neq US,v}$
$\Delta \log(1+\tau_{i,US,v})$	-0.0171	0.00758	-0.0193
	(0.0200)	(0.0108)	(0.0227)
N	4,847,859	1,552,523	355,727
Fixed Effects			
Destination-Variety	✓		
Origin-Destination	✓		
Sample	All	$\Delta\tau \neq 0$	Chinese Origin

Notes: This table estimates the pass-through of US tariffs on scale-free prices to third-party countries. We report first-difference OLS estimates with varying fixed effects and controls for scale economies. We perform a two-stage least squares estimates, first netting out scale economies and then regressing the residual scale-free price on US tariffs. Column (1) includes destination-variety and origin-destination fixed effects. Column (2) only considers products with changes in the US-origin Import tariff rate. Column (3) only considers products originating in China. Standard errors are clustered at the product-origin level. See the text for full details and sources.

residual price $\tilde{p}$. We find no significant effect. This empirical result is extremely tight around zero:

$$\frac{d \log \tilde{c}_{o,US,v}}{d \log(1+\tau_{o,US,v})} \approx 0.$$

Markups To complete our incidence calculation, we require an estimate for the level of markups, $\tilde{\mu}$. We do this two ways. We first compute our own markup estimates using the methodology of Feenstra (1994) as detailed in Appendix B.3. This approach assumes CES demand and monopolistic competition to recover markups from trade data.[31] Second, as the CES may be inconsistent with our markups changes, we consider production-side estimates, drawing from studies that use production data from China which find modal markups around 1.3 for manufacturing exporters (Yue and Lin, 2023). As this production function estimation models put little structure on demand, we consider it a good complement to our own estimates, which rely on the Feenstra (1994) framework. Both approaches yield similar markup estimates. We additionally show that our results are robust to a wide range of markup estimates, from 1.1 to 1.5, which encompasses the range

[31]There is a tension: the CES monopolistic competition model that underlies this method is globally inconsistent with our main findings. Specifically, a standard CES model has constant markups. We employ this method primarily because the Feenstra (1994) estimator identifies the supply and demand elasticity from the variance-covariance structure of prices and quantities, and complements our IV approaches for recovering the supply elasticity. We could extend this approach to include variable markups.

Table 15: Transaction Scale Changes and Tariff Pass-Through

	(1)	(2)	(3)	(4)
	2017-2019 Change Pricing Scale: $\Delta\gamma^{OLS}_{o,d,v}$			
Δ(1+Tariff)	-0.0264	-0.00581	0.181	0.0127
	(0.0567)	(0.0102)	(0.123)	(0.131)
Fixed Effects			Country	HTS6
Weighted	Observations	Quantile Regression	Observations	Observations

Notes: This table reports regressions of the change in the estimated quantity discount parameter ($\Delta\gamma$) between 2017 and 2019 on the change in tariffs over the same period. The dependent variable is the change in the OLS-estimated scale parameter from Table 4 recovered at the level of origin-destination-variety-year. Column (1) is a simple OLS regression weighted by the number of transactions. Column (2) reports a quantile regression at the median. Columns (3) and (4) include country and HTS6 fixed effects, respectively. Robust standard errors are in parentheses. See the text for full details and sources.

of estimates in the literature.

Scale Changes Implicit in our analysis, we have assumed that the scale elasticity $\gamma_{o,v}$ is constant across all products and origins and invariant to tariffs. We test this assumption by estimating changes in $\gamma_{o,v,year}$ directly from the data before and after the 2018 tariffs. We retain the OLS framework from Table 4 and allow γ to vary by origin, variety, and year.[32] Table 15 shows that the changes in the scale elasticity is invariant to tariff changes from 2017 to 2019. Column (1) is the baseline regression in first differences. Columns (2-4) conduct robustness and controls for other factors such as country and HTS6 codes or considering quantile regression.

Welfare Implications Table 16 presents a welfare decomposition of the US-China trade war, under different assumptions. We combine observed changes in trade values and quantities from 2018 to 2019, applied tariff rates, and estimated pass-through and scale elasticities to compute the changes in buyer surplus, foreign producer surplus, and government revenue. Reported figures are then projected on a 2017 baseline and reported on an annualized basis in 2017 dollars. Row 1 shows our baseline estimate: a $22.9 billion reduction in foreign producer surplus versus a $16.6 billion loss for domestic buyers. This implies an incidence ratio (I) of 0.72, very different from non-scale adjusted estimates where buyers bear all the burden.

This finding helps discern the underlying market structure. At the bottom of Table 16,

[32]Using the IV is problematic here, given concerns about first stage power.

we report theoretical benchmarks derived from Section 2. Under a standard model like Perfect Competition, incidence is dominated by domestic buyers (Incidence ratio of -∞ in our case because pass-through exceeds 100 percent). Our baseline result of $I = 0.72$ is incompatible with this standard framework. Instead, it lies closer to the "One-Price Monopoly" benchmark ($I = 1.01$), suggesting an important role for exporters to absorb tariffs by reducing markups.

Comparing the "Baseline" (Row 1) to "Rest of World" (Row 3), we see the mechanism in action. For non-China trade—where tariffs were lower and supply chains less disrupted—the incidence ratio (0.85) is higher. In contrast, for US-China trade (Row 2) the incidence ratio is lower at 0.71. This suggests that when faced with larger tariffs, Chinese exporters reduced their markups more than their counterparts elsewhere.

In aggregate, the tariffs functioned as a tool of international redistribution, but with significant efficiency costs. Our calculation in Row 1 suggests a net gain for the US (Buyer Loss + Govt Revenue $\approx$ \$13.4 billion gain). This result relies on our finding that the relevant price for Buyer Surplus is the scale-free price $\tilde{p}$, not the observed unit value. While observed unit values rose one-for-one with tariffs, a significant portion of this increase was due to a mechanical slide up the cost curve as buyers reduced order sizes. By the Envelope Theorem, the welfare loss from this quantity adjustment is second-order; the primary distortion is the shift in the price schedule itself, which rose by only 60%.

However, this domestic gain came at the cost of a massive \$22.9 billion profit shock to foreign producers. This figure captures both the reduction in markups and the true "efficiency loss" of the trade war—the real resources consumed by processing smaller, more frequent batches. Because transaction-level costs rose as scale economies were lost, foreign sellers faced a double squeeze: lower effective prices and higher per-unit costs, consistent with the production declines documented in Chor and Li (2024).[33] Overall, world welfare declined by $\approx$ \$9.5 billion. By exercising terms-of-trade power via tariffs, the US successfully induced a squeeze on foreign profits, extracting rents as predicted by relevant trade theory (Ossa, 2014; Bagwell and Staiger, 1999), albeit at a large cost to aggregate efficiency. This analysis does not account for any foreign retaliation so we cannot fully outline the full cost to both US and global welfare.

It is important to qualify that the Envelope Theorem provides a first-order approximation of welfare changes. While standard in the literature (Weyl and Fabinger, 2013; Ganapati et al., 2020), our estimated quantity responses suggest that second-order effects ("Harberger triangles") could be non-negligible. Specifically, if the disutility of frequent ordering rises non-linearly, the approximation might understate the administrative bur-

[33] A caveat is that we do not attribute profits of multinationals to ultimate owners.

den on domestic buyers. However, even if we were to treat the loss of scale entirely as a direct cost to buyers, the qualitative failure of the "complete pass-through" result remains, as the reduction in $\tilde{p}$ provides a substantial transfer.

Extensions and Robustness These tariff burden results are robust across a range of alternative specifications. Rows 4-12 of Table 16 show that using different estimates for scale economies yield broadly similar results, with the seller's share of the burden remaining substantial and incidence ratios between 0.6 and 0.85. Row 4 uses OLS-derived γ_v and Row 5 uses the IV-derived common γ from Section 4. Row 6 uses the same mean $\mu = 1.3$ across industries, Row 7 uses estimates of μ from Yue and Lin (2023)'s cost based approach from Chinese microdata, and Row 8 uses $\mu = 1.5$. Row 9 considers the CES price index, instead of the baseline quantity weighting. Row 10 uses the longer-run changes from 2017-2019, instead of the baseline monthly variation (referring to Panel E in Table 12).

The main outlier is the "Transaction-Level Pass-Through" estimate (Row 11), which suggests buyers bore a larger burden (although still a finite incidence ratio!). This is because it is calculated using only intensive-margin adjustments within continuing relationships, thereby missing the crucial effects of firm entry and exit and other compositional shifts that are captured in our aggregate analysis.

One potential concern may be that we use $\gamma_\mu = 0$. To test the quantitative significance of this, we instead assume that scale economies are entirely driven by markups instead (and thus $\gamma_c = 0$ instead). In Row 12, we re-calculate incidence under this assumption and find that sellers would still bear a large share of the burden (Incidence = 0.54).

Our baseline welfare calculations primarily capture small changes. However, as shown in Table 12, the data reveal a significant extensive margin response, with a large change in transaction frequency. In Appendix A.4, we generalize our incidence framework to explicitly account for these extensive margin effects. As of now, we are unable to directly compare our local result to the general-equilibrium result in Fajgelbaum et al. (2020).

However, by locally approximating the surplus lost from exiting transactions in Row 13, we show that the total welfare impact is larger than the intensive-margin estimates suggest and an order of magnitude larger than Fajgelbaum et al. (2020). While the qualitative distribution of the burden remains similar, there is now much larger aggregate efficiency loss, as the extensive margin changes represent a true loss of mutually beneficial trade relationships. In this case, domestic buyer welfare now declines by $30.3 billion, outweighing government revenue. We leave a more precise comparison to Ganapati and Hottman (2026), where we fully estimate a consistent demand system.

Table 16: Welfare Effects of Tariffs

	Specification	Change in Surplus (Billion 2017 USD)				Incidence (Ratio)	Share of Surplus Change	
		Seller	Buyer	Gov.	Total	Buyer / Seller	Seller vs Buyer	Domestic vs Foreign
1	Baseline	-22.9	-16.6	30.0	-9.5	0.72	58%	-141%
2	Baseline - China Only	-20.9	-14.9	26.8	-9.0	0.71	58%	-132%
3	Baseline - Rest of World	-2.0	-1.7	3.2	-0.5	0.85	54%	-300%
4	OLS Scale Discounts	-26.9	-16.6	30.0	-13.5	0.62	62%	-99%
5	Common Scale Discounts	-27.0	-16.6	30.0	-13.6	0.61	62%	-99%
6	Common Markups	-19.7	-16.6	30.0	-6.3	0.84	54%	-213%
7	Exogenous Markups	-22.1	-16.6	30.0	-8.7	0.75	57%	-154%
8	Common High Markups	-24.7	-16.6	30.0	-11.3	0.67	60%	-119%
9	CES Price Index	-21.4	-18.2	30.1	-9.5	0.85	54%	-125%
10	2017-2019 Only	-22.1	-16.6	30.0	-8.7	0.75	57%	-154%
11	Transaction Level Pass-Through	-9.9	-23.7	30.6	-3.0	2.39	30%	-228%
12	Markup Driven Scale	-19.5	-16.6	30.0	-6.1	0.85	54%	-221%
13	Extensive Margin	-35.9	-30.3	30.0	-36.2	0.84	54%	1%
14	Perfect Competition					$-\infty$	-1%	-
15	One-Price Monopoly					1.01	50%	1%
16	Perfectly Discriminatory Monopoly					0.00	100%	$-\infty$

Notes: This table presents the annualized welfare effects of the cumulative change in US tariffs in 2018 and 2019 under various specifications projected onto 2017 trade data. The first section (Rows 1-13) shows our baseline and robustness checks, reporting changes in foreign seller surplus, domestic buyer surplus, government revenue, total welfare change, incidence ratios, and shares of surplus change. Total represents the sum of Seller, Buyer, and Government revenue changes. The second section (Rows 14-16) provides theoretical benchmarks for incidence under different market structures. See text for details on each specification. Markup estimates are derived from either our own calculations using the Feenstra (1994) method, cost-based estimates from Yue and Lin (2023), or assumed values. Scale discount estimates are derived from either our OLS estimates or IV estimates from Section 4. The "Extensive Margin" row (13) incorporates first-order welfare effects from transaction entry and exit, as detailed in Appendix A.4.

7.3 Rationalizing the Pass-Through Mechanism

How do we reconcile our findings of incomplete base-price pass-through ($\tilde{\rho} \approx 0.6$) with the existence of significant scale economies? In a standard competitive model with constant markups, increasing returns to scale (a downward-sloping supply curve) implies that tariff pass-through should exceed 100 percent. Intuitively, because tariffs reduce the quantity demanded, and lower quantities drive up marginal costs (by moving firms up the cost curve), the price increase should be amplified beyond the tariff itself.

To rationalize these findings, we interpret them through a structural framework that accounts for simultaneous quantity discounts, imperfect competition, and variable markups. We again adapt the sufficient statistic approach of Weyl and Fabinger (2013) to our setting

with ad-valorem tariffs, following Adachi and Fabinger (2022).[34] The unit-price pass-through elasticity ρ is determined by the interaction of demand, scale, and conduct:

$$\rho \equiv \frac{d \log P}{d \log(1+\tau)} = \left[\underbrace{\frac{\epsilon_D - \theta}{\epsilon_D}}_{\text{Stabilizer}}\right] \left[1 + \underbrace{\frac{\theta}{\epsilon_{ms}}}_{\text{Curvature}} + \underbrace{\frac{\theta}{\epsilon_\theta}}_{\text{Conduct}} + \underbrace{\frac{\epsilon_D - \theta}{\epsilon_S}}_{\text{Scale Effect}}\right]^{-1} \tag{23}$$

where ϵ_D is the price elasticity of demand, ϵ_S is the supply elasticity (the inverse of the elasticity of marginal cost with respect to quantity), ϵ_{ms} is the inverse elasticity of marginal surplus (capturing demand curvature), and ϵ_θ is the inverse elasticity of the conduct parameter with respect to quantity. The conduct parameter θ ranges from 0 (perfect competition) to 1 (monopoly). See Appendix A.3 for the full derivation.

This decomposition illustrates that observed pass-through is the outcome of two opposing forces.

The Scale Effect vs. Markup Adjustment Counteracting the scale effect is the opposing force of *Markup Adjustment* (which is only possible when conduct is not perfectly competitive so $\theta \neq 0$). First, consider the leading term $\frac{\epsilon_D - \theta}{\epsilon_D}$, which acts as a "stabilizer". As Delipalla and Keen (1992) show, ad-valorem tariffs act as automatic stabilizers compared to specific taxes: because the tax payment scales with the price, firms face a stronger disincentive to raise markups. This stabilizer effect persists even with non-linear pricing. Second, the demand "curvature" and changing "conduct" terms in the denominator capture additional mechanisms through which markup adjustment can dampen pass-through (and these would be the only mechanisms in the case of specific taxes). The conduct term reduces pass-through if $\varepsilon_\theta > 0$, and thus higher tariffs make conduct less competitive. Holding conduct fixed, the demand curvature term reduces pass-through if $\frac{1}{\varepsilon_{ms}} > 0$. This parameter $1/\epsilon_{ms}$—the elasticity of marginal surplus—is analogous to the "super-elasticity" of demand identified by Klenow and Willis (2016) and Kimball (1995) as the source of real rigidities in macroeconomics. Note that $\frac{1}{\varepsilon_{ms}} = 1 - \rho^{MN}$ where ρ^{MN} is the demand curvature term in Mrázová and Neary (2017). For example, in the "Pollak family" of preferences as formulated and estimated by Arkolakis, Costinot, Donaldson, and Rodríguez-Clare (2019), $\rho^{MN} = [(\frac{\sigma+1}{\sigma})\left(\frac{q_{it}}{q_{it}+\alpha}\right)]$. They estimate $\alpha > 0$, which is necessary for $\frac{1}{\varepsilon_{ms}} > 0$. Economically, $\frac{1}{\varepsilon_{ms}} > 0$ satisfies "Marshall's Second Law of Demand": as tariffs push prices up, the demand elasticity rises, forcing firms to compress markups to defend market share.

[34] We assume that γ is invariant to tariffs, as shown in the empirical exercise. The appendix generalizes results.

In imperfectly competitive markets, firms typically compress markups in response to taxes. The finding of complete pass-through ($\rho \approx 1$) in recent trade war studies (Fajgelbaum et al., 2020; Cavallo et al., 2021) presents a paradox, as it implies either perfect competition or constant markups (CES monopolistic competition). Our estimate of scale-free pass-through $\tilde{\rho} \approx 0.6$ resolves this paradox, confirming that when the mechanical effects of quantity discounts are removed, the behavioral response of exporters aligns with the theoretical prediction that ad-valorem taxes are under-shifted.

8 Conclusion

This paper shows the empirical relevance of quantity discounts at the transaction level for US imports and their importance for aggregate outcomes. These discounts appear to largely reflect scale economies on the supply-side and not market power. Furthermore, accounting for quantity discounts changes our understanding of 2018–2019 tariff pass-through and incidence. Ultimately, we show that the 'pass-through puzzle' is a problem of composition: when trade collapses, efficiency is lost, and standard price indices and pass-through estimates understate foreign incidence.

Our decomposition shows that foreign exporters are effectively fighting a two-front war against tariffs. These two forces are jointly caused by the tariff-induced quantity decline. On one side, the tariff reduces trade volumes, pushing them up their average cost curves (the loss of scale economies). Our evidence suggests these costs are logistical—the rising per-unit cost of shipping, handling, and processing smaller batches. This "supply chain waste" represents a real efficiency loss, consuming resources that could have otherwise been deployed elsewhere. On the other side, to maintain market share, exporters compress their markups. Consequently, the incidence falls much more heavily on foreign producers than previously recognized. This finding challenges the view that US tariffs were largely paid by domestic consumers. And while the more recent 2025 tariffs are more broad-based than those considered in this paper, recent estimates (e.g., Gopinath and Neiman (2026)) continue to suggest 100 percent pass-through of these more recent tariffs to US import prices. Our results show that this result is not sufficient to infer incidence, and indeed if the same mechanisms remain at work, then foreigners could be bearing a significant share of these new tariffs despite the pass-through result.

Our analysis focuses on the direct, partial equilibrium effects of tariffs within affected bilateral relationships. A natural question is whether these scale economies were simply transferred to other countries via trade diversion (e.g., Vietnam or Mexico). We leave general equilibrium frameworks (e.g., Fajgelbaum et al. (2020)) for future work.

Beyond this paper, the appropriate consideration of quantity discounts is important in several other contexts. For example, methodologies that rely on material usage as a proxy for undistorted flexible inputs, commonly used to measure distortions in output and labor markets (Yeh, Macaluso, and Hershbein, 2022), may be biased if quantity discounts are ignored. Similarly, in industrial organization, the link between transaction scale and firm scale (Ganapati, 2025) suggests that studies of sectoral welfare must account for this heterogeneity. Finally, in a macroeconomic context, such quantity discounts affect analyses that rely on aggregate input-output tables or assume market structures where upstream prices are exogenous or linear (Baqaee and Farhi, 2020).

References

ADACHI, T. AND M. FABINGER (2022): "Pass-through, welfare, and incidence under imperfect competition," *Journal of Public Economics*, 211, 104589.

ALESSANDRIA, G., J. P. KABOSKI, AND V. MIDRIGAN (2010): "Inventories, lumpy trade, and large devaluations," *American Economic Review*, 100, 2304–2339.

ALLEN, T. (2014): "Information frictions in trade," *Econometrica*, 82, 2041–2083.

ALVIAREZ, V. I., M. FIORETTI, K. KIKKAWA, AND M. MORLACCO (2023): "Two-sided market power in firm-to-firm trade," *NBER Working Paper*.

AMITI, M., O. ITSKHOKI, AND J. KONINGS (2014): "Importers, exporters, and exchange rate disconnect," *American Economic Review*, 104, 1942–1978.

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.

ANGRIST, J. D. (2014): "The perils of peer effects," *Labour Economics*, 30, 98–108.

ANTRAS, P., T. C. FORT, AND F. TINTELNOT (2017): "The margins of global sourcing: Theory and evidence from US firms," *American Economic Review*, 107, 2514–2564.

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

ANTWEILER, W. AND D. TREFLER (2002): "Increasing returns and all that: a view from trade," *American Economic Review*, 92, 93–119.

ARKOLAKIS, C., A. COSTINOT, D. DONALDSON, AND A. RODRÍGUEZ-CLARE (2019): "The elusive pro-competitive effects of trade," *The Review of Economic Studies*, 86, 46–80.

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.

ATKESON, A. AND A. BURSTEIN (2008): "Pricing-to-market, trade costs, and international relative prices," *American Economic Review*, 98, 1998–2031.

BAGWELL, K. AND R. W. STAIGER (1999): "An economic theory of GATT," *American Economic Review*, 89, 215–248.

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 (2025): "The textbook case for industrial policy: Theory meets data," *Journal of Political Economy*, 133, 1527–1573.

BASU, S. AND J. G. FERNALD (1997): "Returns to scale in US production: Estimates and implications," *Journal of political economy*, 105, 249–283.

BAUMOL, W. J. (1952): "The transactions demand for cash: An inventory theoretic approach," *The Quarterly journal of economics*, 66, 545–556.

BERMAN, N., P. MARTIN, AND T. MAYER (2012): "How do different exporters react to exchange rate changes?" *The Quarterly Journal of Economics*, 127, 437–492.

BERNARD, A. B., E. DHYNE, G. MAGERMAN, K. MANOVA, AND A. MOXNES (2022): "The origins of firm heterogeneity: A production network approach," *Journal of Political Economy*, 130, 1765–1804.

BERRY, S. T. AND P. A. HAILE (2021): "Foundations of demand estimation," in *Handbook of industrial organization*, Elsevier, vol. 4, 1–62.

BISHOP, R. L. (1968): "The effects of specific and ad valorem taxes," *The Quarterly Journal of Economics*, 82, 198–218.

BORNSTEIN, G. AND A. PETER (2025): "Nonlinear pricing and misallocation," *American Economic Review*, 115, 3852–3908.

BRODA, C., N. LIMAO, AND D. E. WEINSTEIN (2008): "Optimal tariffs and market power: the evidence," *American Economic Review*, 98, 2032–2065.

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.

BURSTEIN, A. T., J. CRAVINO, AND M. ROJAS (2024): "Input price dispersion across buyers and misallocation," *NBER Working Paper*.

CAVALLO, A., G. GOPINATH, B. NEIMAN, AND J. TANG (2021): "Tariff pass-through at the border and at the store: Evidence from us trade policy," *American Economic Review: Insights*, 3, 19–34.

CHEUNG, F. K. (1998): "Excise taxes on a non-uniform pricing monopoly: ad valorem and unit taxes compared," *Canadian Journal of Economics*, 1192–1203.

CHOR, D. AND B. LI (2024): "Illuminating the effects of the US-China tariff war on China's economy," *Journal of International Economics*, 150, 103926.

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.

DELIPALLA, S. AND M. KEEN (1992): "The comparison between ad valorem and specific taxation under imperfect competition," *Journal of Public Economics*, 49, 351–367.

DUARTE, M., L. MAGNOLFI, D. QUINT, M. SØLVSTEN, AND C. SULLIVAN (2025): "Conduct and Scale Economies: Evaluating Tariffs in the US Automobile Market," *working paper*.

FAJGELBAUM, P., P. GOLDBERG, P. KENNEDY, A. KHANDELWAL, AND D. TAGLIONI (2024): "The US-China trade war and global reallocations," *American Economic Review: Insights*, 6, 295–312.

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," *Working Paper*.

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

FONTAINE, F., J. MARTIN, AND I. MEJEAN (2020): "Price discrimination within and across EMU markets: Evidence from French exporters," *Journal of international economics*, 124, 103300.

GANAPATI, S. (2025): "The modern wholesaler: Global sourcing, domestic distribution, and scale economies," *American Economic Journal: Microeconomics*, 17, 1–40.

GANAPATI, S. AND C. HOTTMAN (2026): "Micro and Macro Pass-Through: Tariffs and Taxes in a Structural Model of Supply and Demand," *Working Paper*.

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," *Working Paper*.

GOLDBERG, P. K. AND R. HELLERSTEIN (2008): "A structural approach to explaining incomplete exchange-rate pass-through and pricing-to-market," *American Economic Review*, 98, 423–429.

GOPINATH, G., O. ITSKHOKI, AND R. RIGOBON (2010): "Currency choice and exchange rate pass-through," *American Economic Review*, 100, 304–336.

GOPINATH, G. AND B. NEIMAN (2026): "The Incidence of Tariffs: Rates and Reality," *working paper*.

HALPERN, L. AND M. KOREN (2007): "Pricing to firm: An analysis of firm-and product-level import prices," *Review of international economics*, 15, 574–591.

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.

HORNOK, C. AND M. KOREN (2015): "Per-shipment costs and the lumpiness of international trade," *Review of Economics and Statistics*, 97, 525–530.

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., V. LUGOVSKYY, AND A. SKIBA (2009): "The trade reducing effects of market power in international shipping," *Journal of Development Economics*, 89, 84–97.

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," *Working Paper*.

JENKIN, F. (1872): "3. on the principles which regulate the incidence of taxes," *Proceedings of the Royal Society of Edinburgh*, 7, 618–631.

KAMAL, F. AND R. MONARCH (2018): "Identifying foreign suppliers in US import data," *Review of International Economics*, 26, 117–139.

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

KIMBALL, M. S. (1995): "The quantitative analytics of the basic neomonetarist model," *Journal of Money, Credit, and Banking*, 27, 1241–1277.

KLENOW, P. J. AND J. L. WILLIS (2016): "Real rigidities and nominal price changes," *Economica*, 83, 443–472.

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

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.

MANOVA, K. AND Z. ZHANG (2012): "Export prices across firms and destinations," *The Quarterly Journal of Economics*, 127, 379–436.

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

MASKIN, E. AND J. RILEY (1984): "Monopoly with incomplete information," *The RAND Journal of Economics*, 15, 171–196.

MELESHCHUK, S. (2017): "Price Discrimination in International Trade: Empirical Evidence and Theory," *Working Paper*.

MELITZ, M. J. (2003): "The impact of trade on intra-industry reallocations and aggregate industry productivity," *Econometrica*, 71, 1695–1725.

MIRAN, S. (2025): "The Inflation Outlook," https://www.federalreserve.gov/newsevents/speech/miran20251215a.htm, remarks by Gov. Stephen I. Miran at the School of International and Public Affairs, Columbia University, New York, New York.

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

MUNSON, C. L., J. JACKSON, ET AL. (2015): "Quantity discounts: An overview and practical guide for buyers and sellers," *Foundations and Trends in Technology, Information and Operations Management*, 8, 1–130.

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

OSSA, R. (2014): "Trade wars and trade talks with data," *American Economic Review*, 104, 4104–4146.

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

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

STIGLER, G. J. (1961): "The economics of information," *Journal of political economy*, 69, 213–225.

STOLE, L. A. (2007): "Price discrimination and competition," *Handbook of industrial organization*, 3, 2221–2299.

U.S. BUREAU OF LABOR STATISTICS (2025): "International Price Program (IPP): Data Sources," `https://www.bls.gov/opub/hom/ipp/data.htm`, accessed: 2025-10-24.

VERBOVEN, F. (2002): "Quality-based price discrimination and tax incidence: evidence from gasoline and diesel cars," *RAND Journal of Economics*, 275–297.

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.

YUE, W. AND Q. LIN (2023): "Export duration and firm markups: evidence from China," *Humanities and Social Sciences Communications*, 10, 1–11.

A Theoretical Derivations and Microfoundations

A.1 Derivation of Buyer Surplus Change

Buyer surplus is defined as the utility $U(q)$ derived from consuming quantity q minus the total expenditure. Given the non-linear unit price schedule $P(q) = \tilde{p}p(q)$, where $\tilde{p}$ is the tariff-inclusive base price level and $p(q)$ is the scale factor (reflecting quantity discounts), the buyer's total expenditure is $E(q) = q \cdot \tilde{p}p(q)$. The buyer chooses q to maximize surplus:

$$DS(\tilde{p}) = \max_{q} \{U(q) - \tilde{p}p(q)q\} \tag{24}$$

To determine the incidence of a tariff change, we first consider the effect of a change in the base price $\tilde{p}$. By the Envelope Theorem, the change in maximized surplus with respect to the parameter $\tilde{p}$ is equal to the partial derivative of the objective function, evaluating q at the optimal choice:

$$\frac{dDS}{d\tilde{p}} = -\frac{\partial E}{\partial \tilde{p}} = -p(q)q \tag{25}$$

This result mirrors Shephard's Lemma but is modified by the scale factor $p(q)$. It implies that the effective welfare loss from a base price increase is proportional to the *discounted* price paid, not the base price. Differentiating with respect to the tariff term $(1+\tau)$ and applying the chain rule yields:

$$\frac{dDS}{d(1+\tau)} = \int_t \frac{dDS_t}{d\tilde{p}_t} \frac{d\tilde{p}_t}{d(1+\tau)} dt = -\int_t q_t p(q_t) \frac{d\tilde{p}_t}{d(1+\tau)} dt \tag{26}$$

This corresponds to Equation ((2)) in the main text (which reports the magnitude of the incidence). If $p(q) = 1$ (uniform pricing), this collapses to the standard result where welfare loss is simply quantity times the change in price.

Simplification with Iso-elastic Pricing If we assume the scale factor takes the iso-elastic form $p(q) = q^{\gamma}$, then the term $p(q)q$ simplifies to $q^{\gamma} \cdot q = q^{1+\gamma}$. The buyer surplus derivative becomes:

$$\frac{dDS}{d(1+\tau)} = -\int_t q_t^{1+\gamma} \frac{d\tilde{p}_t}{d(1+\tau)} dt \tag{27}$$

This can be re-written in terms of observed transaction value. Let $V_t = P(q_t)q_t = \tilde{p}_t q_t^{1+\gamma}$ be the total value of transaction t. We can rewrite the derivative of the level term as

$\frac{d\tilde{p}_t}{d(1+\tau)} = \tilde{p}_t \frac{d \ln \tilde{p}_t}{d(1+\tau)}$. Substituting these in yields:

$$\frac{dDS}{d(1+\tau)} = -\int_t V_t \frac{d \ln \tilde{p}_t}{d(1+\tau)} dt \tag{28}$$

This result implies that if we define pass-through as the percentage change in the scale-free price ($\frac{d \ln \tilde{p}_t}{d \ln(1+\tau)}$), aggregate incidence is simply the pass-through multiplied by total transaction value. This is the expression used in the main text.

Caveat on Large Adjustments It is important to qualify that the Envelope Theorem result relies on the assumption of marginal adjustments. If the inventory or administrative cost function $C(q)$ is convex (e.g., rising sharply as order size approaches zero due to fixed logistical overhead), the linearization implied by the Envelope theorem will understate the welfare loss. The lost surplus from the 'intramarginal' units that are no longer purchased—and the lost efficiency from the units now purchased at inefficiently small scales—represents a second-order deadweight loss (analogous to a Harberger triangle, but potentially larger).

A.1.1 Impact on Incidence: γ is Endogenous

When the shape parameter γ is endogenous, the incidence of the tax becomes heterogeneous across consumers of different sizes. Applying the Envelope Theorem to the consumer's problem, the change in consumer surplus is driven by the direct shift in the price schedule at the chosen quantity q:

$$\frac{dCS(q)}{d \log T} = -\text{Expenditure}(q) \times (\tilde{\rho} + \rho_\gamma \log q) \tag{29}$$

This reveals that the "effective" pass-through for welfare analysis is not a single number but a function of quantity: $\tilde{\rho}_{\text{eff}}(q) = \tilde{\rho} + \rho_\gamma \log q$. Consequently, relying on the aggregate average $\tilde{\rho}$ may mask distributional effects where the tax burden is shifted disproportionately onto large or small buyers depending on the sign of ρ_γ.

A.2 Derivation of Producer Surplus Change

In this section, we derive the change in producer surplus with respect to tariffs as presented in Equation (3).

Producer surplus is defined as the integral of profits over all transactions:

$$PS = \int_{t \in T} \pi(t)\, dt \tag{30}$$

$$= \int_{t \in T} [R(t) - C(t)]\, dt \tag{31}$$

where $R(t)$ denotes revenue and $C(t)$ denotes total variable costs for transaction t. Note that revenue is a function of the tariff-inclusive price faced by the buyer, stripped of the tariff wedge. Let τ be the ad-valorem tariff rate.

The revenue for a single transaction is:

$$R(t) = \frac{P(q_t)}{1+\tau} \cdot q_t = \frac{\tilde{p} \cdot p(q_t)}{1+\tau} \cdot q_t \tag{32}$$

The total cost for a single transaction is:

$$C(t) = \text{Unit Cost}(q_t) \cdot q_t = [\tilde{c} \cdot c(q_t)] \cdot q_t \tag{33}$$

We are interested in the derivative of aggregate producer surplus with respect to the tariff factor $(1+\tau)$. By linearity of the integral, we can differentiate term by term:

$$\frac{dPS}{d(1+\tau)} = \int_t \left[\frac{dR(t)}{d(1+\tau)} - \frac{dC(t)}{d(1+\tau)} \right] dt \tag{34}$$

A.2.1 Revenue Decomposition

Differentiating the logarithm of revenue $R(t)$ with respect to $\log(1+\tau)$:

$$\log R = \log \tilde{p} + \log p(q) + \log q - \log(1+\tau) \tag{35}$$

$$\frac{d \log R}{d \log(1+\tau)} = \frac{d \log \tilde{p}}{d \log(1+\tau)} + \frac{d \log p(q)}{d \log q} \frac{d \log q}{d \log(1+\tau)} + \frac{d \log q}{d \log(1+\tau)} - 1 \tag{36}$$

Grouping terms involving q:

$$\frac{d \log R}{d \log(1+\tau)} = \frac{d \log \tilde{p}}{d \log(1+\tau)} + \left(1 + \frac{d \log p(q)}{d \log q}\right) \frac{d \log q}{d \log(1+\tau)} - 1 \tag{37}$$

Converting from elasticities back to levels using $dX/X = d \log X$:

$$\frac{dR}{d(1+\tau)} = \left[\frac{d \log \tilde{p}}{d \log(1+\tau)} + (1+\gamma) \frac{d \log q}{d \log(1+\tau)} - 1 \right] \frac{R}{1+\tau} \tag{38}$$

where $\gamma = \frac{d \log p(q)}{d \log q}$ is the elasticity of the price schedule (quantity discount).

A.2.2 Cost Decomposition

Similarly, for costs $C(t)$:

$$\log C = \log \tilde{c} + \log c(q) + \log q \tag{39}$$

$$\frac{d \log C}{d \log(1+\tau)} = \frac{d \log \tilde{c}}{d \log(1+\tau)} + \left(1 + \frac{d \log c(q)}{d \log q}\right) \frac{d \log q}{d \log(1+\tau)} \tag{40}$$

Assuming base costs $\tilde{c}$ are constant with respect to tariffs (no input linkages): $\frac{d \log \tilde{c}}{d \log(1+\tau)} = 0$. However, retaining the general form:

$$\frac{dC}{d(1+\tau)} = \left[\frac{d \log \tilde{c}}{d \log(1+\tau)} + (1+\gamma^c) \frac{d \log q}{d \log(1+\tau)}\right] \frac{C}{1+\tau} \tag{41}$$

Substituting these components back into the integral yields the result in the main text.

A.3 Recovering Pass-Through

We follow Weyl and Fabinger (2013), but with both ad-valorem taxes and endogenous pricing (subject to caveats).

A.3.1 Demand and Pricing

Consumers face a non-linear price schedule set by the firm. The unit price p depends on quantity q according to the schedule:

$$p(q) = \tilde{p} \cdot q^{\gamma} \tag{42}$$

where γ is a fixed parameter capturing the non-linearity (e.g., bulk discounts if $\gamma < 0$, surcharges if $\gamma > 0$).

Assumption (Consumer Internalization): The consumer fully observes and internalizes the shape of this schedule (the quantity discount or surcharge determined by γ), but takes the base price parameter $\tilde{p}$ as given. They maximize utility $U(q) - p(q)q$. The First Order Condition ($U'(q) = p(q) + p'(q)q$) yields the inverse demand curve effectively faced by the firm:

$$p(q) = \frac{U'(q)}{1+\gamma} \tag{43}$$

A.3.2 Firm Behavior

The firm chooses the base price $\tilde{p}$ to maximize profit, given the schedule shape γ and a conduct parameter θ (where $\theta = 0$ implies perfect competition and $\theta = 1$ implies monopoly). The firm faces an ad-valorem tax (or tariff) $T = 1 + \tau$. The equilibrium condition is:

$$MR(q) = MC(q) \cdot T \tag{44}$$

Marginal Revenue is defined structurally by conduct θ:

$$MR(q) = p\left(1 - \frac{\theta}{\epsilon_D}\right) \tag{45}$$

A.3.3 Elasticity Definitions

To derive pass-through, we define the following structural elasticities:

- **Demand Elasticity (ϵ_D):** Defined on the inverse demand curve $p(q)$ faced by the firm:
$$\frac{1}{\epsilon_D} \equiv -\frac{dp}{dq}\frac{q}{p}$$
- **Marginal Surplus (ms):** Defined as the absolute gap between price and marginal revenue for a monopolist (or the slope of the inverse demand times quantity). It represents the consumer's inframarginal surplus on the last unit:
$$ms(q) \equiv -p'(q)q = \frac{p(q)}{\epsilon_D}$$
Note that we can rewrite Marginal Revenue as $MR(q) = p(q) - \theta ms(q)$.
- **Curvature of Marginal Surplus (ϵ_{ms}):** The **inverse elasticity** of the marginal surplus function, which captures the curvature of demand (often related to the convexity of the demand curve), consistent with Weyl and Fabinger (2013):
$$\epsilon_{ms} \equiv ms/(\frac{dms}{dq}q)$$
- **Conduct Elasticity (ϵ_θ):** While θ is often treated as a parameter, it may vary with quantity (e.g., if collusion stability depends on volume). We define its **inverse elasticity** as:
$$\epsilon_\theta \equiv \theta/(\frac{d\theta}{dq}q)$$

A.3.4 Derivation of Unit Price Pass-Through (ρ)

We seek the pass-through rate of the tax to the observable unit price, defined as $\rho \equiv \frac{d \log p}{d \log T}$.

A.3.5 Log-Linearization

Taking logs of the equilibrium condition $MR = MC \cdot T$:

$$\log MR(q) = \log MC(q) + \log T \tag{46}$$

Differentiating with respect to $\log q$ and multiplying by the response of quantity to tax $\frac{d \log q}{d \log T}$:

$$\frac{d \log MR}{d \log q} \frac{d \log q}{d \log T} = \frac{d \log MC}{d \log q} \frac{d \log q}{d \log T} + 1 \tag{47}$$

Rearranging to solve for the quantity response:

$$\frac{d \log q}{d \log T} = \frac{1}{\frac{d \log MR}{d \log q} - \frac{d \log MC}{d \log q}} \tag{48}$$

By definition of demand elasticity, the price change is linked to quantity by $\frac{d \log p}{d \log q} = -\frac{1}{\epsilon_D}$. Thus $\rho = -\frac{1}{\epsilon_D} \frac{d \log q}{d \log T}$. Substituting this in:

$$\rho = \frac{-1/\epsilon_D}{\frac{d \log MR}{d \log q} - \frac{1}{\epsilon_S}} = \frac{1}{-\epsilon_D \frac{d \log MR}{d \log q} + \frac{\epsilon_D}{\epsilon_S}} \tag{49}$$

A.3.6 Expansion of Marginal Revenue

We expand the term $\frac{d \log MR}{d \log q}$. Using $MR = p - \theta ms$ and differentiating with respect to q:

$$\begin{aligned} \frac{dMR}{dq} &= p' - (\theta' ms + \theta ms') \\ &= p' - \theta ms \left(\frac{\theta'}{\theta} + \frac{ms'}{ms} \right) \end{aligned}$$

Converting to elasticities. Recall $p' = -p/(q\epsilon_D)$ and $ms = p/\epsilon_D$:

$$\frac{dMR}{dq} = \frac{dp}{dq} \left[1 + \theta \frac{ms}{-p'} \left(\frac{1}{q\epsilon_\theta} + \frac{1}{q\epsilon_{ms}} \right) \right] \tag{50}$$

Note that $\frac{ms}{-p'} = \frac{-p'q}{-p'} = q$. Thus:

$$\frac{dMR}{dq} = \frac{dp}{dq}\left[1 + \theta(\epsilon_\theta^{-1} + \epsilon_{ms}^{-1})\right] \tag{51}$$

Finally, converting to the logarithmic derivative $\frac{d\log MR}{d\log q} = \frac{dMR}{dq}\frac{q}{MR}$:

$$\frac{d\log MR}{d\log q} = \underbrace{\left(\frac{dp}{dq}\frac{q}{p}\right)}_{-\frac{1}{\epsilon_D}}\frac{p}{MR}[1 + \theta\epsilon_{ms}^{-1} + \theta\epsilon_\theta^{-1}] \tag{52}$$

A.3.7 Pass-Through

Substituting this back into Eq ((49)), and using the identity $p/MR = \epsilon_D/(\epsilon_D - \theta)$, we substitute $\frac{p}{MR}$ and factor it out to obtain the structural pass-through formula:

$$\rho = \frac{1}{\frac{\epsilon_D}{\epsilon_D - \theta}\left[1 + \theta/\epsilon_{ms} + \theta/\epsilon_\theta + \frac{\epsilon_D - \theta}{\epsilon_S}\right]} \tag{53}$$

Note: This factorization highlights that pass-through is scaled by the inverse markup. The formula depends on γ only implicitly (as γ affects the equilibrium values of ϵ_D and ϵ_S), but the structural form is identical to the standard case.

Comparison to the Specific-Tax Formulation This is a scaled version of the specific tax formulation in Weyl and Fabinger (2013). In particular:

$$\rho = \rho_{\text{ST}} \cdot \frac{mr}{p} = \rho_{\text{ST}} \cdot \frac{\epsilon_D - \theta}{\epsilon_D}.$$

Where the ad-valorem pass-through $\rho = \frac{d\log p}{d\log T}$ is linked to the specific tax pass-through $\rho_{\text{ST}} = \frac{dp}{dt}$ via an "equivalent" specific tax shock.

There are a few differences. The elasticity formulation means we do not use the Lerner markup rule, but rather just the internalized marginal revenue term and do not worry about the specific tax starting at $t = 0$. The relationship between specific tax t and ad-valorem tax $1 + \tau$ is exactly inverse of the markup.

Let $t(T)$ be a specific tax that yields the same equilibrium marginal cost shift as the ad-valorem tax T at the current optimum:

$$t(T) = (T - 1)MC$$

Differentiating this effective burden with respect to T gives the intensity of the cost shock:

$$\frac{dt}{dT} = MC$$

Using the chain rule, the price response to the ad-valorem tax is the specific pass-through rate times this cost shock intensity:

$$\frac{dp}{dT} = \frac{dp}{dt}\frac{dt}{dT} = \rho_{\text{ST}} \cdot MC$$

To convert this to an elasticity, we multiply by $\frac{T}{p}$:

$$\rho = \frac{T}{p}\frac{dp}{dT} = \frac{T}{p}(\rho_{\text{ST}} \cdot MC) = \rho_{\text{ST}}\frac{MC \cdot T}{p}$$

Using the equilibrium condition $MR = MC \cdot T$, we substitute the numerator:

$$\rho = \rho_{\text{ST}}\frac{MR}{p}$$

A.3.8 Base Price Pass-Through ($\tilde{\rho}$)

While ρ measures the change in the realized unit price, welfare analysis requires tracking the shift in the pricing *schedule*. We define the base price pass-through as:

$$\tilde{\rho} \equiv \frac{d\log\tilde{p}}{d\log T} \tag{54}$$

Relationship between ρ and $\tilde{\rho}$ From the pricing schedule $p = \tilde{p}q^{\gamma}$, we differentiate with respect to $\log T$:

$$\rho = \tilde{\rho} + \gamma\frac{d\log q}{d\log T} = \tilde{\rho} + \gamma(-\epsilon_D\rho) \tag{55}$$

Solving for $\tilde{\rho}$:

$$\tilde{\rho} = \rho(1 + \gamma\epsilon_D) \tag{56}$$

This geometric link holds regardless of market structure or conduct. It relates the movement *along* the schedule (determined by ρ) to the shift *of* the schedule ($\tilde{\rho}$).

A.4 Generalizing the Framework: The Extensive Margin

Our local incidence analysis in Section 2.2 captures the intensive margin (changes in transaction size) but omits the extensive margin (changes in transaction frequency). We now extend the framework to incorporate this margin and provide a complete welfare calculation.

The change in producer surplus, $dPS = dR - dC$, can be calculated using our empirical findings. Since quantity discounts are cost-driven ($\gamma = \gamma_c$) and scale-free costs are unaffected by tariffs ($d\ln\tilde{c}/d\ln(1+\tau) = 0$), we can relate costs to revenues via the markup: $C = R/\tilde{\mu}$. The change in producer surplus is then:

$$\frac{dPS}{d(1+\tau)} = \frac{R}{1+\tau}\left[\frac{d\ln R}{d\ln(1+\tau)}\left(1-\frac{1}{\tilde{\mu}}\right)+\frac{1}{\tilde{\mu}}\frac{d\ln\tilde{\mu}}{d\ln(1+\tau)}\right]. \tag{57}$$

All terms on the right-hand side are either directly estimated from our tariff regressions or taken from the literature, allowing us to compute the incidence on producers without direct cost data.

The downstream buyer surplus calculation in Equation (2) is incomplete, as it only captures the welfare change for continuing transactions (the intensive margin). Our results show a strong extensive margin response, with a significant drop in the number of transactions. A more complete expression for the change in downstream buyer surplus is:

$$\frac{dDS}{d(1+\tau)} = \int_{t\in\mathcal{T}_{cont}} \frac{d\tilde{p}_t}{d(1+\tau)} q_t dt - \int_{t\in\mathcal{T}_{exit}} DS_t dt, \tag{58}$$

where $\mathcal{T}_{cont}$ is the set of continuing transactions, $\mathcal{T}_{exit}$ is the set of exiting transactions, and DS_t is the surplus from a transaction prior to its exit. While precisely calculating the lost surplus from exiting transactions (DS_t) is difficult without a full demand model.

A simple approximation is to assume that the average surplus lost per exiting transaction is proportional to the average surplus lost per continuing transaction. If we further assume that the pre-tariff quantity of exiting transactions is similar to that of continuing ones, we can approximate the total lost surplus from the extensive margin by scaling the intensive margin loss by the relative number of exiting to continuing transactions. This leads to the following approximation for the total change in downstream buyer surplus:

$$\frac{dDS}{d(1+\tau)} = \frac{d(T\cdot\overline{DS})}{d(1+\tau)} = \frac{dT}{d(1+\tau)}\overline{DS} + T\frac{d\overline{DS}}{d(1+\tau)}, \tag{59}$$

where T is the number of transactions and $\overline{DS}$ is the average downstream buyer surplus per transaction. The first term captures the extensive margin (change in the number of

transactions), while the second term captures the intensive margin (change in surplus per transaction). We can approximate the change in average surplus using our intensive margin estimate for continuing transactions: $T\frac{d\overline{DS}}{d(1+\tau)} \approx \int_{t\in\mathcal{T}_{cont}} \frac{d\tilde{p}_t}{d(1+\tau)} q_t dt$.

A.5 Microfoundation for IV Strategy

Our IV strategy exploits two procurement structures. Buyers make monthly inventory decisions that allocate total demand across suppliers. Buyers also then split each supplier's monthly quantity across multiple transactions for logistical reasons. These levels generate different correlation patterns and correspond to our two IV specifications.

A.5.1 Inventory and Supplier Allocation

Buyer j faces monthly demand $D_{j,t} = \bar{D}_j \cdot \exp(\varepsilon_{j,t})$ where $\varepsilon_{j,t}$ is a monthly shock. Following inventory models (Arrow et al., 1951), the buyer orders to maintain target stock:

$$Q_{j,v,t} = \kappa \cdot D_{j,t} - I_{j,t-1} \approx \kappa \bar{D}_j \cdot \varepsilon_{j,t} \tag{60}$$

This total is allocated across suppliers $i \in \mathcal{I}_j$ to minimize cost. With quantity discounts $p_i(Q) = \tilde{p}_i Q^\gamma$ and fixed transaction costs A_i, the optimal allocation is:

$$Q_{i\to j,v,t} = \text{share}_i \cdot Q_{j,v,t} \quad \text{where} \quad \text{share}_i = \frac{\tilde{p}_i^{-1/\gamma}}{\sum_k \tilde{p}_k^{-1/\gamma}} \tag{61}$$

Purchases from different suppliers move together in a positive correlation:

$$\log Q_{i\to j,v,t} = \log(\text{share}_i \kappa \bar{D}_j) + \varepsilon_{j,t} \tag{62}$$

A.5.2 Transaction Allocation Within Supplier-Month

Each supplier's monthly total $Q_{i\to j,v,t}$ is split across transactions $\{q_{i\to j,\tau,v}\}_{\tau\in t}$ based on logistics (shipping schedules, inventory cycles, payment terms). These factors are orthogonal to transaction-specific pricing shocks ξ_τ (quality issues, rush delivery, spot pricing).

The budget constraint is:

$$\sum_{\tau\in t} q_{i\to j,\tau,v} \approx Q_{i\to j,v,t} \tag{63}$$

Transactions within the same month are mechanically related in a negative correlation:

$$q_{i\to j,\tau,v} \approx Q_{i\to j,v,t} - \sum_{\tau'\neq\tau} q_{i\to j,\tau',v} \tag{64}$$

A.5.3 Two IV Strategies

Within-Month Allocation (Column 3 - Baseline).

$$IV^{(2)}_{i\to j,\tau,v} = \log \sum_{\tau'\neq\tau,\tau'\in t} q_{i\to j,\tau',v} \tag{65}$$

Relevance: From (63), $IV^{(2)} \approx Q_{i\to j,v,t} - q_\tau \to$ negative correlation
Exclusion: $\mathbb{E}[\xi_{i,\tau}|IV^{(2)}] = 0$ if transaction shocks don't affect:

(i) Monthly procurement plan $Q_{i\to j,v,t}$ (predetermined by Level 1)

(ii) Allocation timing of other transactions (determined by logistics)

Cross-Supplier (Column 6).

$$IV^{(1)}_{i\to j,v,t} = \log \sum_{k\neq i} Q_{k\to j,v,t} \tag{66}$$

Relevance: Both Q_i and $IV^{(1)}$ depend on $\varepsilon_{j,t} \to$ positive correlation
Exclusion: $\mathbb{E}[\xi_{i,\tau}|IV^{(1)}] = 0$ if transaction shocks to supplier i don't affect:

(i) Total monthly demand $Q_{j,v,t}$ (predetermined by downstream production)

(ii) Allocation to other suppliers (determined by base prices $\{\tilde{p}_k\}$)

A.5.4 Role of Fixed Effects

Seller-month-variety FE ($\gamma_{i,v,t}$) control for supplier i's total activity in month t, absorbing common demand shocks across buyers. This leaves identification from within-month allocation, explaining the negative first-stage coefficient in our baseline specification.

Buyer-seller-variety FE ($\mu_{i\to j,v}$) control for time-invariant relationship characteristics (base prices $\tilde{p}_i$, allocation shares).

A.5.5 Reconciling the Estimates

The two IVs yield different estimates ($\gamma = -0.284$ vs -0.200).

Sample selection: Column 3 requires multiple transactions per month; Column 6 requires multiple suppliers. Different selections may capture different LATEs if γ is heterogeneous.

Weak IV bias: Column 6's weaker first stage ($F = 115$) may induce finite-sample bias toward OLS (-0.268), making the estimate smaller in magnitude.

Exclusion violations: If either exclusion restriction is violated, estimates diverge. The similarity suggests both are approximately valid.

We report Column 6 ($\gamma = -0.200$) as the conservative estimate for robustness checks.

A.5.6 Relation to Existing Literature

This framework combines inventory models (Arrow et al., 1951; Alessandria et al., 2010) with "leave-one-out" IV strategies (Angrist, 2014). The negative first-stage coefficient does not invalidate the IV—mechanical negative correlation from budget constraints is standard in peer effects and network literatures when group totals are fixed.

B Additional Empirical Results and Robustness Checks

B.1 Comparison of Scale Across Methods

Table A.1 compares the variety-level scale elasticity estimates (γ) from Section 4.2. Each observation is an HTS 6-digit variety. The table reports regressions of the IV estimates (γ^{IV}), structural estimates ($\gamma^{Structural}$), and shipping cost-based estimates ($\gamma^{Shipping}$) on the baseline OLS estimates (γ^{OLS}). The structural estimates are computed using a method similar to Feenstra (1994), which assumes locally iso-elastic demand. Columns (1)-(4) report unweighted regressions, while columns (5)-(8) report regressions weighted by the number of observations in each variety. The results show a strong positive correlation between the different estimation methods, particularly between the OLS and IV estimates.

We take these results as evidence that the scale elasticity estimates are robust across different methods and that the OLS and IV estimates are reliable indicators of the true scale elasticity. The strong correlation between these estimates suggests that the differences in the methods used to estimate γ do not significantly alter the overall findings.

In particular, the cost-based estimates are particularly interesting as they provide a different perspective on the scale elasticity, which is not directly influenced by the demand

structure or the availability of data on downstream buyer behavior. The strong correlation between the shipping cost-based estimates and the other methods further supports the robustness of the scale elasticity estimates.

Table A.1: Comparison of Scale Elasticity Estimates (γ) Across Varieties

	(1) γ^{IV}	(2) $\gamma^{Structural}$	(3) $\gamma^{Structural}$	(4) $\gamma^{Shipping}$	(5) γ^{IV}	(6) $\gamma^{Structural}$	(7) $\gamma^{Structural}$	(8) $\gamma^{Shipping}$
γ^{OLS}	0.834		0.782	0.652	0.906		1.167	1.157
	(0.00974)		(0.0251)	(0.0214)	(0.00764)		(0.0249)	(0.0141)
γ^{IV}		0.683				0.975		
		(0.0244)				(0.0256)		
R^2	0.631	0.155	0.185	0.173	0.766	0.254	0.339	0.602
Weighting	Unweighted				Weighted - Observations			

Notes: This table compares the variety-level scale elasticity estimates (γ) from Section 4.2. Each observation is an HTS 6-digit variety. The table reports regressions of the IV estimates (γ^{IV}), structural estimates ($\gamma^{Structural}$), and shipping cost-based estimates ($\gamma^{Shipping}$) on the baseline OLS estimates (γ^{OLS}). The structural estimates are computed using a method similar to Feenstra (1994), which assumes locally iso-elastic demand. Columns (1)-(4) report unweighted regressions, while columns (5)-(8) report regressions weighted by the number of observations in each variety.

B.2 Recovered Scale Elasticity Correlates

In Table A.2, we perform a falsification test of the price discrimination hypothesis. We relax the assumption of a common scale elasticity and instead recover a unique scale elasticity $\gamma_{i,j,v}$ for every buyer-seller pair using the strategy from Section 4.2. If quantity discounts were driven by second-degree price discrimination, we would expect the steepness of the discount schedule to vary with the relative market power of the trading partners.

We regress these relationship-specific elasticities on proxies for market power, including related party status, bilateral volume, and market shares. Standard errors are clustered at the variety level. As shown in Table A.2, we find precise null results: none of the market power proxies are statistically significant predictors of $\gamma_{i,j,v}$. This lack of correlation is strong evidence against the price discrimination mechanism. It supports our main interpretation that γ reflects a technological parameter (e.g., logistical cost structure) that is common across transactions, rather than a strategic variable adjusted for specific relationships.

In A.3, we explore the correlates of year-on-year changes in variety scale elasticities from 2016 to 2017. Each observation is a variety scale elasticity change for a buyer-seller

Table A.2: Correlates of Buyer-Seller Scale Elasticities

	(1) $\gamma_{i,j,v}$	(2) $\gamma_{i,j,v}$	(3) $\gamma_{i,j,v}$	(4) $\gamma_{i,j,v}$	(5) $\gamma_{i,j,v}$	(6) $\gamma_{i,j,v}$
Related Party	-0.756					
	(0.717)					
Bilateral Quantity: $\log(q_{ij})$		0.00856				
		(0.0964)				
Output Share			-0.159			
			(0.773)			
Input Share				0.0905		
				(0.505)		
Quantity Sold: $\log(q_i)$					-0.00435	
					(0.0144)	
Quantity Bought: $\log(q_j)$						0.00993
						(0.00852)
R^2	0.25	0.25	0.25	0.25	0.0187	0.0187
Within R^2	0.0000101	8.45E-09	5.09E-08	3.34E-08	2.77E-08	0.00000018
Fixed Effects	Buyer-Variety, Seller-Variety				Variety	

Notes: This table uses the OLS strategy from Section 4.2 to recover a scale-elasticity for every buyer-seller pair in the data. Each observation is a buyer-seller-variety scale elasticity. Standard errors are clustered at the variety level.

pair. We regress the change in scale elasticities on changes in various factors such as changes in HHI for both buyers and sellers, changes in the number of buyers, sellers, pairs, transactions, and value. Standard errors are clustered at the variety level. We do not find that the changes in scale elasticity are significantly correlated with any of these factors.

B.3 Structural Estimation of Supply and Demand Elasticities

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

Table A.3: Correlates of Changes in Buyer-Seller Scale Elasticities

	(1) $\Delta\gamma^{OLS}$	(2) $\Delta\gamma^{OLS}$	(3) $\Delta\gamma^{OLS}$	(4) $\Delta\gamma^{OLS}$	(5) $\Delta\gamma^{OLS}$	(6) $\Delta\gamma^{OLS}$	(7) $\Delta\gamma^{OLS}$
Δ HHI Seller	0.639						
	(0.520)						
Δ HHI Buyer		0.359					
		(0.203)					
$\Delta \log(\text{Buyers})$			0.31				
			(0.275)				
$\Delta \log(\text{Sellers})$				0.162			
				(0.194)			
$\Delta \log(\text{Pairs})$					0.149		
					(0.149)		
$\Delta \log(\text{Transactions})$						0.00444	
						(0.0316)	
$\Delta \log(\text{Value})$							-0.00412
							(0.0101)
R^2		0.003	0.001	0.003	0.001	0.001	0.000

Notes: This table explores the correlates of year-on-year changes in buyer-seller scale elasticities (γ) from 2016 to 2017. Each observation is a buyer-seller-variety scale elasticity. We regress the change in scale elasticities on changes in various factors. Standard errors are clustered at the variety level.

As an alternative, we can put structure on demand (say CES), and we can jointly estimate supply and demand, following Feenstra (1994), and in an application similar to Hottman, Redding, and Weinstein (2016). There is one downside to this, all demand and supply functions must have fully defined functional forms. 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 supply side. This assumption does need a parametrization 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

$$\Delta^{r,t} \ln(q_{vt}) = (-\sigma^s)\Delta^{r,t} \ln(p_{vt}) + \nu_{vt}, \tag{67}$$

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

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

$$\Delta^{r,t} \ln p_{vt} = \omega^s \Delta^{r,t} \ln(q_{vt}) + \kappa_{vt}, \tag{68}$$

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

$$G(\beta_s) = \mathbb{E}_{\mathbb{T}} \left[\nu_{vt}\kappa_{vt}(\beta_s)\right] = 0 \tag{69}$$

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:

$$\mathbb{E}_{\mathbb{T}} \left[(\Delta^{r,t} \ln p_{vt})^2\right] = \mathbb{E}_{\mathbb{T}} \left[(\frac{\omega^s}{\sigma^s})(\Delta^{r,t} \ln q_{vt})^2 + (\frac{\omega^s\sigma^s - 1}{\sigma^s})\Delta^{r,t} \ln q_{vt}\Delta^{r,t} \ln p_{vt} + \frac{1}{\sigma^s}\nu_{vt}\kappa_{vt}\right], \tag{70}$$

which simplifies to:

$$\mathbb{E}_{\mathbb{T}} \left[(\Delta^{r,t} \ln p_{vt})^2\right] = (\frac{\omega^s}{\sigma^s})\mathbb{E}_{\mathbb{T}} \left[(\Delta^{r,t} \ln q_{vt})^2\right] + (\frac{\omega^s\sigma^s - 1}{\sigma^s})\mathbb{E}_{\mathbb{T}} \left[\Delta^{r,t} \ln q_{vt}\Delta^{r,t} \ln p_{vt}\right]. \tag{71}$$

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.[35] 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 two adaptations. First, instead of considering products at the root 10-digit level, we aggregate to the 4-digit level. Second, instead of using the raw transactions at the daily level, we aggregate to the monthly level. Overall, we are left with 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 largely 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 4.2, furthermore it requires a grid search to account for multiple solutions. Following the literature, we do primarily 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 to be 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, where $\omega \approx \gamma_v \approx -0.3$.

Table A.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

Notes: See text for details.

Table A.4 shows our results. Fundamentally, ω is similar to the γ estimated, but with

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

a different strategy. We consider this below.

We first choose the solution with $\sigma > 0$ and with the smallest inverse supply elasticity. In this main specification (Specification 1), we find a median supply elasticity of -0.233 and a mean elasticity of -0.365, 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 logarithmic and exponential) transformation, finds a mean demand elasticity of approximately of $\sigma = 3.4$, which is broadly in line with the literature.

Overall, this is broadly consistent with the supply elasticities estimated in the more reduced-form exercise Section 4.2, as well as firm-demand estimates of σ from Hottman et al. (2016).[36]

B.4 Alternative Time Periods

In table A.5, we replicate Table 11 for the year 2019. We use the same specifications as in Table 11, but we only include transactions from 2019. The results show that the pass-through of tariffs to observed prices is still significant, but the magnitude of the effect is lower compared to the 2017-2018 period. However the gap between the effect for raw unit values (column 1) and the effect for quantity-adjusted values (columns 2 and 3) is similar compared to the 2018 tariffs.

Table A.6 reports the results of estimating the aggregate pass-through of tariffs to observed unit values, quantity-adjusted unit values, and other relevant variables for the period 2017-2018. We drop the 2019 tariffs, which did not have long to be in before the COVID-19 shock. The table includes fixed effects at the product-origin-month level and clusters standard errors at the product-origin level. The coefficients in column (1) represent the reduced-form pass-through of tariffs to observed unit values, while column (2) reports pass-through to quantity-adjusted unit values, constructed by netting out transaction-level scale effects using the estimated γ from Section 4.2. Column (3) decomposes the scale-adjusted price into the composite of markups and marginal costs $\mu\tilde{c}$, isolating the change in seller pricing behavior net of the mechanical tariff effect. Column

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

Table A.5: Transaction-level Tariff Pass-Through: Within Relationships: 2019 Only

	(1) $\log(p)$	(2) $\log(\tilde{p}^{OLS})$	(3) $\log(\tilde{p}^{IV})$	(4) $\log(pq)$	(5) $\log(q)$	(6) $\log(p)$
log(1 + Tariffs Applied)	0.865	0.667	0.648	-0.0563	-0.921	
	(0.0284)	(0.0212)	(0.0366)	(0.0324)	(0.0432)	
log(1+ Tariffs Statutory)						0.975
						(0.340)
R^2	0.953	0.956	0.991	0.715	0.897	0.954
Within R^2	0.00104	0.00105	0.000522	0.00000271	0.00049	0.00000124
Fixed Effects	Buyer-Seller-Variety, Variety-Year-Month					

Notes: This table reports transaction-level tariff pass-through estimates for continuing buyer-seller-variety relationships in 2019. All specifications include buyer-seller-variety and variety-year-month fixed effects. Standard errors are clustered at the relationship level. Column (1) shows the pass-through of applied tariffs (duty paid) to observed transaction prices. Columns (2) and (3) report pass-through to quantity-adjusted prices ($\tilde{p}$), constructed using the OLS and IV estimates of the scale elasticity γ from Section 4.2, respectively. These columns isolate the change in the price schedule from compositional effects. Column (4) reports the effect on transaction value, while Column (5) shows the response of transaction quantity. Column (6) shows the pass-through of statutory tariffs.

(4) reports the residual quantity composition effect $\tilde{q}$. Column (5) shows the effect on total transaction values pq, column (6) reports the quantity response, column (7) shows the change in average transaction size q/T, and column (8) reports the change in the number of transactions T. Column (9) presents the first-order approximation of scale-adjusted pass-through computed as $-\gamma\times$ (coefficient from column 7), using $\gamma = -0.29$ from Section 4.2.

B.5 Decomposition of Aggregate Price Variation

We can run regressions, along the vein of the decomposition in Section 3, to understand what correlates to aggregate relationship-level prices, as those studied by Alviarez et al. (2023) and Kamal and Sundaram (2016):

$$\log \tilde{p}_{i\to j,v} = \theta \log q_{i\to j,v} + FE_{i,v} + FE_{j,v} + \epsilon_{ij}. \tag{72}$$

Table A.7 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 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 quan-

Table A.6: Monthly Aggregate Pass-Through, 2017-2018

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	$\log p$	$\log \tilde{p}$	$\log \mu\tilde{c}$	$\log \tilde{q}$	$\log pq$	$\log q$	$\log(q/T)$	$\log T$	$\widehat{\log \tilde{p}}$
log(1+Tariffs)	1.01	0.593	-0.417	0.427	-1.493	-2.502	-1.881	-0.621	0.540
	(0.0376)	(0.0149)	(0.0149)	(0.0280)	(0.0395)	(0.0548)	(0.0472)	(0.0229)	
r^2	0.908	0.878	0.878	0.902	0.87	0.9	0.892	0.904	
r^2 within	0.0003	0.0006	0.0006	0.0001	0.0006	0.0009	0.0007	0.0003	
Fixed Effects	Variety-Year-Month, Variety-Country Origin								

Notes: This table reports aggregate tariff pass-through estimates at the product-origin-month level over 2017-2018. Column (1) shows the reduced-form pass-through of tariffs to observed unit values. Column (2) reports pass-through to quantity-adjusted unit values $\tilde{p}$, constructed by netting out transaction-level scale effects using the estimated γ from Section 4.2. Column (3) decomposes the scale-adjusted price into the composite of markups and marginal costs $\mu\tilde{c}$, isolating the change in seller pricing behavior net of the mechanical tariff effect. Column (4) reports the residual quantity composition effect $\tilde{q}$. Column (5) shows the effect on total transaction values pq. Column (6) reports the quantity response. Column (7) shows the change in average transaction size q/T. Column (8) reports the change in the number of transactions T. Column (9) presents the first-order approximation of scale-adjusted pass-through computed as $-\gamma\times$ (coefficient from column 7), using $\gamma = 0.29$ from Section 4.2. All specifications include product-time and product-origin fixed effects. Standard errors are clustered at the product-origin level.

tities 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) include the most demanding set of fixed effects, buyers-variety and seller-variety. Column (3) shows that there is a relatively significant correlation between raw prices and aggregate quantities, with θ = -.137, explaining 13.5% of the variation. However, when controlling for transaction level quantities in (4), there is an extremely small effect, with almost no relationship between the quantity bought and the price paid.[37]

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 better, we turn to a first difference strategy in the main text, 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.

[37]We can additionally run regressions on importer and exporter shares and volumes, as in section 3.

Table A.7: Relationship-Level Cross Sectional Price Differences

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

Notes: See text for details.

C Domestic Trade Data

C.1 CFS Data Descriptions

For domestic trade, we consider the U.S. Census and Department of Transportation Commodity Flow Survey (CFS), focusing on 2012. This survey asks 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 (thus a buyer is a product-location combination). Second, for goods 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.

This survey of approximately 100,000 establishments (out of 700,000 potential establishments) covers locations that originate 80% of domestic manufactured good shipping. Only a subset of transactions are recorded to keep reporting requirements tractable, resulting in about 5 million transactions in 2012. We decompose prices for 2012, but carry out further analysis every 5 years from 1997 to 2017.

Table A.8: Variance Decomposition of 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 Variety		37.6%			62.4%
3		Seller-Buyer-Variety		73.7%		26.4%
4	$\log q$	Seller-Buyer-Variety	18.6%	47.3%	12.0%	22.1%
5	$\log q$ x Variety	Seller-Buyer-Variety	20.4%	45.6%	12.4%	21.6%

Notes: We decompose 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 is fixed to remain constant over the sample period. See the text for full details and specification.

C.1.1 Decomposing Domestic Trade Prices

We conduct a similar exercise with the US domestic Commodity Flow Survey (CFS). The data are limited in certain aspects, particularly because they represent a highly restricted

sample from the buyer's perspective.[38] We broadly find similar trends to the international trade import data.[39]

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

In Table A.8, Row (1) replicates the exercise to decompose the variance of prices on the size of a transaction (all demeaned and residualized at the SCTG 5-digit level). Transactions are identified with a seller, product code, shipment date and buying zip code tuple.

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 is indeed a bit lower than those in the international trade data, however that data has better measures of product 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 the buyer-seller pair, or if there is any explanatory power due to transaction-level quantities. Broadly, as with international trade, the results are similar.

C.2 Supply Estimation: Domestic Data

We repeat a version of the supply isolation using the domestic shipment data from the CFS. The available data make 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. Our primary identification threat is unobserved qualities across shipments.

As before, we create an instrumental variable that is correlated with a demand shock,

[38] This is important since many types of production function estimation assume away price variation in input data.

[39] This analysis is replicable in public use data that aggregates regions and SCTG codes.

but uncorrelated with the current supplier (for which we have implicitly a supplier-time control). Following Equation 10, we create three instruments. The first looks at how much of a product is bought from a firm by a destination state, excluding the focal order. 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 A.9 produces a version of Table 2-3 for our domestic data in 2012. Column (1) illustrates in the simply demeaned data a correlation, where γ is equivalent to -0.3; broadly in line with the import 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 across 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 our first instrument. Column (6)-(8) show that all three instrumentation strategies yield 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 goods traded are different and product level comparisons may yield even more consistent estimates.

C.2.1 Aggregate Scale Economies 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 once quantity discounts are controlled for. 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 and 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 A.10, columns (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

Table A.9: 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		✓		✓

	(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	✓	✓	✓	✓
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 refers to process of regularizing all variables by product code (SCTG) 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. See the text for full details and sources.

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, scale is $-.11$. However, controlling for transaction level quantity discounts, scale is $= .02$, close to constant returns to scale.

Table A.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					✓	✓

Notes: This comes from a regression of change in aggregates quantities on a measure of price. Odd columns use aggregate prices. Even columns follow adjust each transaction's price for the quantity sold, using the aggregate relationship using γ to adjust p to recover the scale-free $\tilde{p}$. 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 5-digit product. See the text for full details and sources.