ONLINE APPENDIX

A Results with Non-homothetic CES Framework

A.1 Non-homothetic CES Preference and Price Index

We introduce non-homotheticities using the non-separable class of CES functions in Sato (1975), Comin, Lashkari and Mestieri (2021), Matsuyama (2019) and Redding and Weinstein (2020), which satisfy implicit additivity in Hanoch (1975). The non-homothetic CES consumption index for item i, C^m_{ib}, is defined by the following implicit function:

$$\sum_{k\in m}\left(\frac{\varphi^m_{kib}C^m_{kib}}{(C^m_{ib})^{(\epsilon_{kib}-\sigma_{ib})/(1-\sigma_{ib})}}\right)^{\frac{\sigma_{ib}-1}{\sigma_{ib}}}=1 \tag{11}$$

where C^m_{kib} denotes total consumption of barcode k; σ_{ib} is the elasticity of substitution between barcodes; ϵ_{kib} is the constant elasticity of consumption of barcode k with respect to the consumption index (C^m_{ib}) which controls the income elasticity of demand for that barcode. Assuming that barcodes are substitutes ($\sigma_{ib}>1$), it is required $\epsilon_{kib}<\sigma_{ib}$ for the consumption index to be globally monotonically increasing and quasi-concave, and therefore to correspond to a well-defined utility function. When $\epsilon_{kib}=1$ for all $k\in m$, the utility function becomes homothetic.

We solve the expenditure minimization problem for a given barcode within an item and basic heading to obtain the following expressions for the price index (P^m_{ib}) dual to the consumption index (C_{kib}) and the expenditure share for a individual barcode k (s^m_{kib}):

$$P^m_{ib}=\left(\sum_{k\in m}(p^m_{kib}/\varphi^m_{kib})^{1-\sigma_{ib}}(C^m_{ib})^{\epsilon_{kib}-1}\right)^{\frac{1}{1-\sigma_{ib}}} \tag{12}$$

$$s^m_{kib}=\frac{(p^m_{kib}/\varphi^m_{kib})^{1-\sigma_{ib}}(C^m_{ib})^{\epsilon_{kib}-1}}{\sum_{l\in m}(p^m_{lib}/\varphi^m_{lib})^{1-\sigma_{ib}}(C^m_{lib})^{\epsilon_{lib}-1}}=\frac{(p^m_{kib}/\varphi^m_{kib})^{1-\sigma_{ib}}(E^m_{ib}/P^m_{ib})^{\epsilon_{kib}-1}}{(P^m_{ib})^{1-\sigma_{ib}}} \tag{13}$$

Taking ratios of the shares of Mexico and the United States and rearranging, we obtain the following expression for the difference in the cost of living, which holds for each common barcode available in two countries (common):

$$\frac{P^m_{ib}}{P^u_{ib}}=\frac{p^m_{kib}/\varphi^m_{kib}}{p^u_{kib}/\varphi^u_{kib}}\left(\frac{E^m_{ib}/P^m_{ib}}{E^u_{ib}/P^u_{ib}}\right)^{\frac{\epsilon_{kib}-1}{1-\sigma_{ib}}}\left(\frac{s^m_{kib}}{s^u_{kib}}\right)^{\frac{1}{\sigma_{ib}-1}},\quad k\in\text{common} \tag{14}$$

Summing expenditures across common barcodes, we obtain the following expression for the aggregate share of common barcodes in total expenditure for Mexico and the United

States (λ_{ib}^m and λ_{ib}^u):

$$\lambda_{ib}^m \equiv \frac{\sum_{k\in\text{common}} p_{kib}^m C_{kib}^m}{\sum_{k\in m} p_{kib}^m C_{kib}^m} \quad \text{and} \quad \lambda_{ib}^u \equiv \frac{\sum_{k\in\text{common}} p_{kib}^u C_{kib}^u}{\sum_{k\in u} p_{kib}^u C_{kib}^u} \tag{15}$$

Using this expression, the share of an individual barcode in total expenditure (s_{kib}^m) in equation (13) can be re-written as its share of expenditure on common barcodes (s_{kibt}^{m*}) times this aggregate share of common barcodes in total expenditure (λ_{ib}^m):

$$s_{kib}^m = \lambda_{ib}^m s_{kib}^{m*}, \quad k \in \text{common} \tag{16}$$

Taking logs to equation (14) and using equation (16), we obtain the following equation:

$$\log\left(\frac{P_{ib}^m}{P_{ib}^u}\right)^{1+\frac{\epsilon_{kib}-1}{1-\sigma_{ib}}} = \log\frac{p_{kib}^m/\varphi_{kib}^m}{p_{kib}^u/\varphi_{kib}^u} + \log\left(\frac{E_{ib}^m}{E_{ib}^u}\right)^{\frac{\epsilon_{kib}-1}{1-\sigma_{ib}}} + \log\left(\frac{s_{kib}^{m*}}{s_{kib}^{u*}}\right)^{\frac{1}{\sigma_{ib}-1}} + \log\left(\frac{\lambda_{ib}^m}{\lambda_{ib}^u}\right)^{\frac{1}{\sigma_{ib}-1}} \tag{17}$$

We define the ideal log-difference weights (ω_{kib}), the logarithmic mean of common variety expenditure shares, as follows:

$$\omega_{kib}^M = \frac{\dfrac{s_{kib}^{m*} - s_{kib}^{u*}}{\ln s_{kib}^{m*} - \ln s_{kib}^{u*}}}{\sum_{k\in\text{common}} \dfrac{s_{kib}^{m*} - s_{kib}^{u*}}{\ln s_{kib}^{m*} - \ln s_{kib}^{u*}}} \tag{18}$$

where

$$s_{kib}^{m*} = \frac{p_{kib}^m C_{kib}^m}{\sum_{k\in\text{common}} p_{kib}^m C_{kib}^m} \quad \text{and} \quad s_{kib}^{u*} = \frac{p_{kib}^u C_{kib}^u}{\sum_{k\in\text{common}} p_{kib}^u C_{kib}^u}$$

We introduce an assumption that tastes are the same between two countries for each common barcode ($\varphi_{kib}^m = \varphi_{kib}^u$ for all $k \in$ common).

By multiplying the ideal log-difference weights (ω_{kib}) to equation (17), under the assumption that tastes are the same between two countries for each common barcode ($\varphi_{kib}^m = \varphi_{kib}^u$ for all $k \in$ common), we can obtain the non-homothetic CES price index for item i by taking the arithmetic mean across common barcodes (Ω_{it}) and exponents on both sides:

$$\text{Non-homothetic Price Index}_{ib} \equiv \frac{P_{ib}^m}{P_{ib}^u} = \left(\prod_{k\in\text{common}} \left(\frac{p_{kib}^m}{p_{kib}^u}\right)^{\omega_{kib}} \times \left(\frac{\lambda_{ib}^m}{\lambda_{ib}^u}\right)^{\frac{1}{\sigma_{ib}-1}}\right)^{\frac{1}{1-\theta_{ib}}} \left(\frac{E_{ib}^m}{E_{ib}^u}\right)^{\frac{\theta_{ib}}{\theta_{ib}-1}} \tag{19}$$

where

$$\theta_{ib} \equiv \sum_{k \in \text{common}} \omega_{kib} \frac{\epsilon_{kib} - 1}{\sigma_{ib} - 1}$$

and the ratio of λ_{ib}^{m} and λ_{ib}^{u} represents the conventional variety correction term that accounts for different sets of goods available in two countries as in Feenstra (1994) and Broda and Weinstein (2006, 2010).

Define I_C as the set of items with common barcodes. Then, an *aggregate* exact price index can be defined as:

$$\textit{Aggregate}\text{ Non-homothetic Price Index} = \prod_{b} \left(\prod_{i \in I_C} \mathbb{NH}_{ib}^{\omega_{ib}^{*}} \times \left(\frac{\lambda_{b}^{m}}{\lambda_{b}^{u}} \right)^{\frac{1}{\eta_b - 1}} \right)^{\omega_b} \tag{20}$$

where ω_{ib}^{*} and ω_b are the logarithmic mean of the expenditure shares of each item and each basic heading, respectively. The spending shares are defined as:

$$\lambda_{b}^{m} \equiv \frac{\sum_{i \in I_C} p_{ib}^{m} C_{ib}^{m}}{\sum_{i} p_{ib}^{m} C_{ib}^{m}} \quad \text{and} \quad \lambda_{b}^{u} \equiv \frac{\sum_{i \in I_C} p_{ib}^{u} C_{ib}^{u}}{\sum_{i} p_{ib}^{u} C_{ib}^{u}}$$

A.2 Decomposition of Non-homothetic CES Price Index

The non-homothetic CES price index at the basic heading level can be written as a function of the price index developed by the ICP. The relationship between the two indexes can be written as:

$$\frac{\text{Non-homothetic Index}_b}{\mathbb{ICP}_b} = \text{Imputation}_b \times \text{Sampling}_b \times \text{Quality}_b \times \text{Engel-curve Variety}_b \tag{21}$$

where imputation bias, sampling bias, and quality bias are the same as the homothetic case. We define "Engel curve variety bias" to replace "variety bias" in equation 9 from homothetic case.

Engel Curve Variety Bias: This bias measures both the cross-country differences in availability of barcodes and the differences in real consumption across countries and is defined as follows:

$$\text{Engel Curve Variety Bias}_i = \text{Variety Bias}_{ib} \times \left(\frac{E_{ib}^{m} / E_{ib}^{u}}{\mathbb{EPI}_{ib}} \right)^{\frac{\theta_{ib}}{\theta_{ib} - 1}} \tag{22}$$

where

$$\text{Variety Bias}_{ib} \equiv \left(\frac{\lambda_{ib}^{m}}{\lambda_{ib}^{u}} \right)^{\frac{1}{\sigma_{ib} - 1}}. \tag{23}$$

We called the second term Engel curve adjustment since it captures the differences in real consumption across the two countries and depends on the income elasticity of demand of the common barcodes across the two countries. When not all barcodes are common across countries, this term serves as an adjustment to the standard variety bias since the share of common barcodes across countries naturally depends on their income differences. However, even if all barcodes across the two countries are common, the Engel Curve adjustment corrects the price index for the relative importance of each barcode as the relative income of the countries change. If the elasticities of consumption of each barcode with respect to the consumption index equal to one, we are back to the homothetic preferences case. In this case, the Engel-curve variety bias becomes the variety bias as homothetic case.

A.3 Parameter Estimation

Taking estimates of the elasticity of substitution as given, we estimate the constant elasticity of consumption (ϵ_{kib}) for each barcode k with respect to the consumption index (C^m_{kib}) as in Comin, Lashkari and Mestieri (2021):

$$\ln \frac{s^h_{kibt}}{s^h_{\mathbf{K}ibt}} - (1-\sigma_{ib}) \ln \frac{p^h_{kibt}}{p^h_{\mathbf{K}ibt}} = (\epsilon_{kib} - 1) \left(\ln \frac{E^h_{ibt}}{p^h_{\mathbf{K}ibt}} + \frac{1}{(1-\sigma_{ib})} \ln s^h_{\mathbf{K}ibt} \right) + \psi^h_t + \epsilon^h_{kibt} \tag{24}$$

where $\mathbf{K}$ is the benchmark barcode, which corresponds to the largest selling barcode in each item, and ψ^h_t is the set of fixed effects. We aggregate households into seven groups by their annual household income. With the US Nielsen data, equation 24 is estimated with quarter×Census region fixed effects. Note that because barcodes are substitutes within all items ($\sigma_{ib} > 1$), it is required $\epsilon_{kib} < \sigma_{ib}$ for the consumption index to be globally monotonically increasing and quasi-concave, and therefore to correspond to a well-defined utility function.[19]

[19] In less than one percent of the cases, $\epsilon_{kib} >= \sigma_{ib}$. In these cases, we impute $\epsilon_{kib} = \sigma_{ib} - 0.01$.

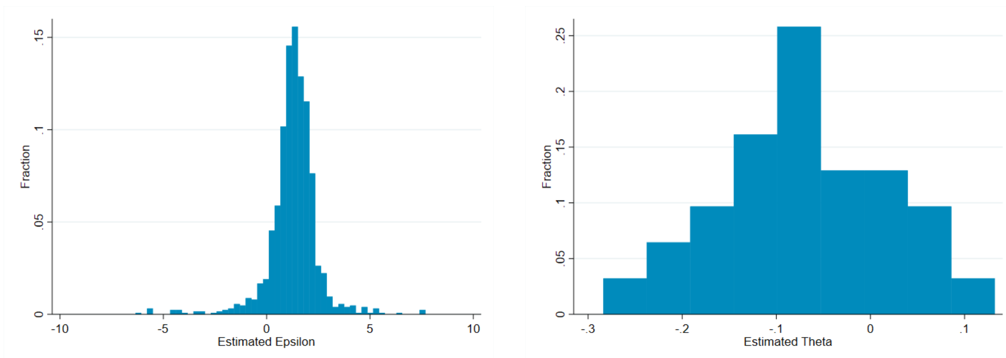
Table A.I: Descriptive Statistics of Estimated Parameters

	Mean	Std. Dev.	10th-Percentile	Median	90th-Percentile
σ_{ib}	6.56	3.13	4.33	5.55	10.53
ϵ_{kib}	1.29	1.21	0.22	1.37	2.31
θ_{ib}	-0.07	0.09	-0.18	-0.08	0.04
E_{ib}^m/E_{ib}^u	0.81	0.69	0.23	0.65	1.60

Note: This table reports descriptive statistics for the elasticity of substitution (σ_{ib}), the elasticity of consumption of barcode k with respect to the consumption index (ϵ_{kib}), parameter in the non-homothetic CES price index (θ_{ib}) and nominal expenditure ratio (E_{ib}^m/E_{ib}^u).

The first row of Table A.I reports descriptive statistics for σ_{ib}. The average of the elasticity of substitution we use is 6.56 with standard deviation of 3.13. The second row of Table A.I reports the descriptive statistics for ϵ_{kib}. Our estimates for this parameter have a mean of 1.29 and a standard deviation of 1.21. The third row of Table A.I reports the descriptive statistics for θ_{ib}. Recall that this parameter is the equally weighted average of $\frac{\epsilon_{kib}-1}{\sigma_{ib}-1}$ across the common barcodes within an item across the two countries. Our estimates of this parameter have a mean of -0.07 and a standard deviation of 0.09. Note that when θ_{ib} is close to zero, the expenditure ratio plays a small role in the price index. Lastly, we report other informative moments for the quantification of the Engel-curve variety bias such as the nominal expenditure ratio. As the last row of Table A.I shows, the ratio varies across items. It has a median of 0.65 and a mean of 0.81 (with standard deviation of 0.69), which indicates that the distribution is skewed to the right. Figure A.1 shows the distribution of estimated ϵ_{kib} and θ_{ib}. The mean of ϵ_{kib} is close to one. As a result, the mean of θ_{ib} is around zero.

Figure A.1: Distribution of Estimated ϵ_{kib} and θ_{ib}



ϵ_{ki} θ_i

Note: Panel (a) and (b) show the distribution of estimated ϵ_{kib} and θ_{ib}, respectively.

A.4 Decomposition Results

Equation 21 indicates that the gap between the two price indexes can be decomposed into the imputation bias, the sampling bias, the quality bias, and the quality-variety bias. The aggregate bias is estimated to be 0.76, which is very close to the homothetic preference case (0.77). This is mainly because, for the common barcodes, the average elasticity of consumption is estimated to be close to 1.

Table A.II: Decomposition Results for the Non-Homothetic CES Price Index

Aggregate Price Index		Bias due to:			Aggregate
Exact	Pseudo-ICP	Imputation	Sampling	Quality-Variety	Bias
0.65	0.86	0.89	0.87	0.99	0.76

Notes: The table reports the aggregate non-homothetic CES price index, pseudo ICP price index, and the gap between the exact and ICP index due to imputation, sampling, and quality-variety. Aggregate bias is the product of the bias due to imputation, sampling, and quality-variety.

B Logit Specification of Barcode-level Price Aggregation

In the main text, we use Cobb-Douglas aggregation as the first-order approximation of CES aggregation (Kmenta, 1967). In this section, we show that CES aggregation can be derived from the aggregation of the choices of individual consumers with extreme-value-distributed idiosyncratic amenities (net of costs) from different stores.

Changing the notion of McFadden (1974) slightly, we suppose that the utility of an individual consumer i who consumes q_{is} units of barcode k from store s is given by:

$$U_i = u_s + a_{is}, \quad u_s \equiv \ln q_{is} \tag{25}$$

where a_{is} captures idiosyncratic amenities to visit store s that are drawn from an independent Type-I Extreme Value distribution:

$$G(a) = e^{-e^{(-a/\nu+\kappa)}} \tag{26}$$

where ν is the scale parameter of the extreme value distribution and κ is the Euler-Mascheroni constant.

Each consumer has the same expenditure on barcode k, E_k, and chooses their preferred store given the observed realization for idiosyncratic amenities. Therefore, the consumer's budget constraint implies:

$$q_{is} = \frac{E_k}{p_s} \tag{27}$$

The probability that individual i choose store s is:

$$\begin{aligned} x_{ikt} &= \text{Prob}[u_{is} + c_{is} > u_{il} + c_{il}, \forall l \neq s] \\ &= \text{Prob}[c_{il} < c_{is} + u_{is} - u_{il}, \forall l \neq s] \end{aligned} \tag{28}$$

Using the distribution of idiosyncratic amenities in equation 26, we have:

$$x_{is}|a_{is} = \prod_{l \neq k} e^{-e^{-(a_{is}+u_{is}-u_{il})/\nu+\kappa}} \tag{29}$$

Once we integrate it across the probability density function for a_{is} and use the change of variable technique as in Anderson, De Palma and Thisse (1992), the probability that individual i chooses store s becomes:

$$s_{is} = s_s = \frac{p_s^{-1/\nu}}{\sum_{l \in \Psi} p_l^{-1/\nu}} \tag{30}$$

The expected utility of consumer i is:

$$E[U_i] = E[\max\{u_{i1} + a_{i1}+, ..., u_{iN} + a_{iN}\}] = \nu \ln \left[\sum_{l \in \Psi} \exp\left(\frac{u_{il}}{\nu}\right) \right] \tag{31}$$

Using the definition of u_{is} in equation 25 and the budget constraint in equation 27, expected utility can be written as:

$$E[U_i] = E_k/P \tag{32}$$

where P is the unit expenditure function:

$$P = \left[\sum_{s \in \Psi} p_s^{-1/\nu} \right]^{-\nu} \tag{33}$$

Therefore, theoretically consistent aggregation of prices of a specific barcode across stores is the CES aggregation.

C Parameter Estimation

In order to obtain the elasticity of substitution, σ_{ib}, for each item, we rely on the method developed by Feenstra (1994) and extended by Broda and Weinstein (2006) and Broda and Weinstein (2010). The procedure consists of estimating a demand and supply equation for each barcode by using only the information on prices and quantities. For this estimation, we face the standard endogeneity problem for a given barcode. Although we cannot identify supply and demand, the data do provide information about the joint distribution of supply and demand parameters.

We first model the supply and demand conditions for each barcode within an item. Specifically, we estimate the demand elasticities by using the following system of differenced demand and supply equations as in Broda and Weinstein (2006):

$$\Delta^{\underline{k},t}\ln S_{kibt} = (1-\sigma_{ib})\Delta^{\underline{k},t}\ln P_{kibt} + \iota_{kibt} \tag{34}$$

$$\Delta^{\underline{k},t}\ln P_{kibt} = \frac{\delta_{ib}}{1+\delta_{ib}}\Delta^{\underline{k},t}\ln S_{kibt} + \kappa_{kibt} \tag{35}$$

Note that when the inverse supply elasticity is zero (i.e. δ_{ib}=0), the supply curve is horizontal and there is no simultaneity bias in σ_g. Equations 34 and 35 are the demand and supply equations of barcode k in an item i differenced with respect to a benchmark barcode in the same item. The k^{th} good corresponds to the largest selling barcode in each item. The k-differencing removes any item level shocks from the data.

The identification strategy relies on two important assumptions. First, we assume that ι_{kibt} and κ_{kibt}, the double-differenced demand and supply shocks, are uncorrelated (i.e., $\mathbb{E}_t(\iota_{kibt}\kappa_{kibt}) = 0$). This expectation defines a rectangular hyperbola in (δ_{ib}, σ_{ib}) space for each barcode within an item, which places bounds on the demand and supply elasticities. Because we already removed any item level shocks, we are left with within item variation that is likely to render independence of the barcode-level demand and supply shocks within an item. Second, we assume that σ_{ib} and ω_{ib} are restricted to be the same over time and for all barcodes in a given item.

To take advantage of these assumptions, we define a set of moment conditions for each item i in a basic heading b as below:

$$G(\beta_{ib}) = E_T[\nu_{kibt}(\beta_{ib})] = 0 \tag{36}$$

where $\beta_{ib} = [\sigma_{ib}, \delta_{ib}]'$ and $\nu_{kibt} = \iota_{kibt}\kappa_{kibt}$.

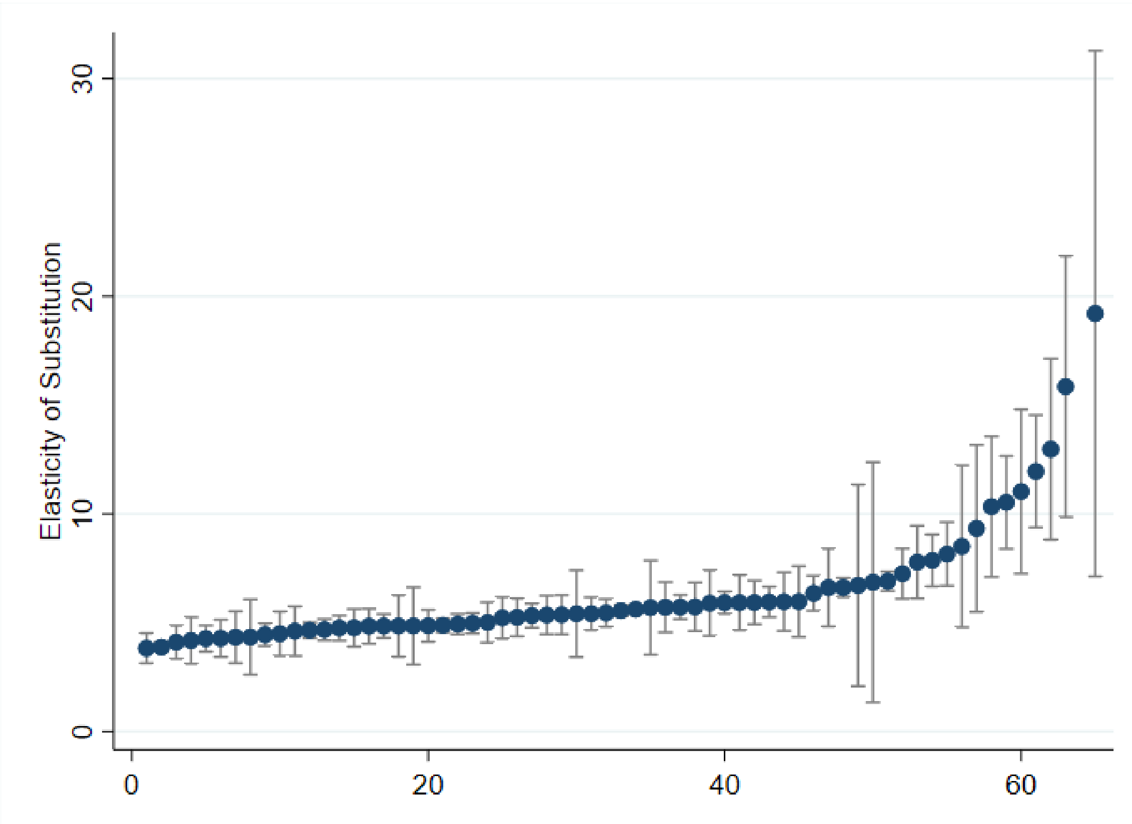
For each item i, all the moment conditions that enter the GMM objective function can be combined to obtain Hansen (1982)'s estimator:

$$\hat{\beta_{ib}} = \arg\min_{\beta_{ib} \in B} G^*(\beta_{ib})' W G^*(\beta_{ib}) \quad \forall\, i \in \omega_b \tag{37}$$

where $G^*(\beta_{ib})$ is the sample analog of $G(\beta_{ib})$, W is a positive definite weighting matrix, and B is the set of economically feasible β_{ib} (i.e., $\sigma_{ib} > 0$). Our estimation procedure follows Redding and Weinstein (2020) using he Nielsen Homescan data from 2004-2019. The elasticities are estimated using data at the quarterly frequency. Households are aggregated using sampling weights to make the sample representative of each country's population. We weight the data for each barcode by the number of raw buyers to ensure that our objective function is more sensitive to barcodes purchased by larger numbers of consumers. We consider barcodes with more 10 or more observations during the estimation. If the procedure renders imaginary estimates or estimates of the wrong sign, we use a grid search to evaluate the GMM objective function above. The average of the elasticity of substitution we obtain is 6.56 with standard deviation of 3.13.[20] Figure C.1 shows the distribution of elasticities sorted by their magnitude and the 95% confidence interval of the point estimates.

[20]Unlike Redding and Weinstein (2020), Broda and Weinstein (2010) and Hottman, Redding and Weinstein (2016) include an additional brand/firm-level layer within a product category. In order to check the robustness of our estimates, we estimate elasticity of substitution within an item-firm level and compare it to our elasticity of substitution within an item level. We use the GS1 data to identify a firm for every barcode. The estimated elasticities of substitution within an item-firm have slightly higher mean (6.91), but both are highly correlated. Correlation is 0.65 and statistically significant at the 1 percent level. Therefore, using elasticity of substitution within an item-firm does not quantitatively affect our results.

Figure C.1: Elasticity of Substitution (σ_{ib})



Note: The figure reports the estimated elasticity of substitution for each item, sorted by its magnitude. The gray lines indicate the 95% confidence intervals.

D Decomposition Results with Common Elasticity of Substitution

In this section, we report decomposition results with common elasticity of substitution across items, $\sigma_{ib} = 6$, which is chosen from a mean of σ_{ib} (6.56). Table D.I reports decomposition results for the exact price index. The aggregate bias (0.75) is estimated to be slightly lower than the item-specific elasticity of substitution case (0.77).

Table D.I: Decomposition Results for the Exact Price Index under $\sigma_{ib} = 6$

Aggregate Price Index		Bias due to:			Aggregate
Exact	Pseudo-ICP	Imputation	Sampling	Quality-Variety	Bias
0.64	0.86	0.89	0.87	0.99	0.75

Notes: The table reports the aggregate exact price index, pseudo ICP price index, and the gap between the exact and ICP index due to imputation, sampling, and quality-variety under $\sigma_{ib} = 6$. Aggregate bias is the product of the bias due to imputation, sampling, and quality-variety.

E Sampling Bias

E.1 Proofs of Propositions

Proposition 1. If the number of basic headings $N_b \to \infty$, the second term of the sampling bias is larger than 1 if $\text{cov}(\omega_{\mathbf{b}}, \ln(\bar{\mathbf{p}}^{\mathbf{c}}_{\mathbf{b}})) > \text{cov}(\omega_{\mathbf{b}}, \ln(\bar{\mathbf{p}}^{\mathbf{u}}_{\mathbf{b}}))$.

Proof. The bias on the second term for country c is:

$$\prod_i \frac{(\bar{p}^c_{ib})^{\omega_{ib}}}{(\bar{p}^c_{ib})^{\frac{1}{N_b}}}$$

This ratio is greater than one if and only if:

$$\sum_i \left(\omega_{ib} - \frac{1}{N_b}\right) \ln(\bar{p}^c_{ib}) > 0 \tag{38}$$

We want to show that this term is equivalent to $\text{cov}(\omega_{\mathbf{b}}, \ln(\bar{\mathbf{p}}^{\mathbf{c}}_{\mathbf{b}}))$ where $\omega_{\mathbf{b}}$ is a vector of weights in basic heading b and $\ln(\bar{\mathbf{p}}^{\mathbf{c}}_{\mathbf{b}})$ is the vector of log prices. By definition:

$$\text{cov}(\omega_{\mathbf{b}}, \ln(\bar{\mathbf{p}}^{\mathbf{c}}_{\mathbf{b}})) = \lim_{N_b \to \infty} \frac{1}{N_b - 1} \sum_i \left(\omega_{ib} - \frac{1}{N_b}\right) \left(\ln(\bar{p}^c_{ib}) - \frac{1}{N_b - 1} \sum_i \ln(\bar{p}^c_{ib})\right)$$

Using that

$$\lim_{N_b \to \infty} \frac{1}{N_b - 1} \sum_i \left(\omega_{ib} - \frac{1}{N_b}\right) \left(\frac{1}{N_b - 1} \sum_i \ln(\bar{p}^c_{ib})\right) = 0$$

Then $\text{cov}(\omega_{\mathbf{b}}, \ln(\bar{\mathbf{p}}^{\mathbf{c}}_{\mathbf{b}}))$ is equivalent to $\frac{1}{N_b - 1} \sum_i \left(\omega_{ib} - \frac{1}{N_b}\right) \ln(\bar{p}^c_{ib})$.

Proposition 2. If the number of stores $S \to \infty$, the third term of the sampling bias is larger than 1 if $\text{cov}(\phi^{\mathbf{m}}_{\mathbf{ib}}, \ln(\bar{\mathbf{p}}^{\mathbf{m}}_{\mathbf{ib}})) - \text{cov}(\phi^{\mathbf{u}}_{\mathbf{ib}}, \ln(\bar{\mathbf{p}}^{\mathbf{u}}_{\mathbf{ib}})) > \text{cov}(\phi^{\mathbf{u}}, \ln(\bar{\mathbf{p}}^{\mathbf{u}}_{\mathbf{ib}})) - \text{cov}(\phi^{\mathbf{m}}, \ln(\bar{\mathbf{p}}^{\mathbf{m}}_{\mathbf{ib}}))$

Proof. Before aggregating across items, the third term of the sample bias for a country c is:

$$\frac{\hat{p}_{ib}^{c}}{\bar{p}_{ib}^{c}} = \frac{\prod_{s\in\Psi^c} \left(\bar{p}_{sib}^{c}\right)^{\phi_{sib}^{c}}}{\prod_{s\in\Psi^c} \left(\bar{p}_{sib}^{c}\right)^{\phi_{s}^{c}}}$$

where ϕ_s^c are total sales weights at the store level and ϕ_{sib}^c are total sales weights at the item-store level. This ratio is larger than 1 for country c if

$$\sum_{s\in\Psi^c} \phi_{sib}^c \ln\left(\bar{p}_{sib}^c\right) > \sum_{s\in\Psi^c} \phi_{s}^c \ln\left(\bar{p}_{sib}^c\right) \tag{39}$$

Let $\phi^{\mathbf{c}}_{\mathbf{ib}}$ be the vector of expenditure weights that vary by item and basic heading, (i.e. $\{\phi_{sib}^c\}_{s\in\Psi^c}$), $\phi^{\mathbf{c}}$ be the vector of weights that only vary at the store level (i.e. $\{\phi_s^c\}_{s\in\Psi^c}$), and $(\bar{\mathbf{p}}^{\mathbf{c}}_{\mathbf{ib}})$ be the vector of log prices (i.e. $\{\ln\left(\bar{p}_{sib}^c\right)\}_{s\in\Psi^c}$). Then equation 39 can be written as:

$$\lim_{S\to\infty} \frac{1}{S-1} \sum_{s\in\Psi^c} \phi_{sib}^c \ln\left(\bar{p}_{sib}^c\right) > \lim_{S\to\infty} \frac{1}{S-1} \sum_{s\in\Psi^c} \phi_{s}^c \ln\left(\bar{p}_{sib}^c\right)$$

Using the definition of covariance on both sides

$$\begin{aligned} &\lim_{S\to\infty} \frac{1}{S-1} \sum_{s\in\Psi^c} \phi_{sib}^c \times \frac{1}{S-1} \sum_{s\in\Psi^c} \ln\left(\bar{p}_{sib}^c\right) + \frac{1}{S-1} \sum_{s\in\Psi^c} \left(\phi_{sib}^c - \bar{\phi}_{ib}^c\right) \left(\ln\left(\bar{p}_{sib}^c\right) - \overline{\ln\left(\bar{p}_{sib}^c\right)}\right) > \\ &\quad \lim_{S\to\infty} \frac{1}{S-1} \sum_{s\in\Psi^c} \phi_{s}^c \times \frac{1}{S-1} \sum_{s\in\Psi^c} \ln\left(\bar{p}_{sib}^c\right) + \frac{1}{S-1} \sum_{s\in\Psi^c} \left(\phi_{s}^c - \bar{\phi}^c\right) \left(\ln\left(\bar{p}_{sib}^c\right) - \overline{\ln\left(\bar{p}_{sib}^c\right)}\right) \end{aligned}$$

Taking the limit as $\lim_{S\to\infty}$ on both sides, we find that $\text{cov}(\phi^{\mathbf{c}}_{\mathbf{ib}}, \bar{\mathbf{p}}^{\mathbf{c}}_{\mathbf{ib}}) > \text{cov}(\phi^{\mathbf{c}}, \bar{\mathbf{p}}^{\mathbf{c}}_{\mathbf{ib}})$.

E.2 Empirical Tests

To quantify the importance of the second term of the sampling bias, we rely on Proposition 1 and test whether $\text{cov}(\omega_{\mathbf{b}}, \ln(\bar{\mathbf{p}}^{\mathbf{m}}_{\mathbf{b}})) > \text{cov}(\omega_{\mathbf{b}}, \ln(\bar{\mathbf{p}}^{\mathbf{u}}_{\mathbf{b}}))$. To do so we rely on the following specification:

$$\omega_{ib} = \alpha + \beta\ \ln(\bar{p}_{ib}^c) \times \mathbb{1}\{c = \text{Mexico}\} + \lambda^c + \theta_b + \epsilon_{ib}^c$$

where the dependent variable is the Sato-Vartia weight for each item. The coefficient of interest is β which indicates whether there is a difference between the covariance between the weights and the prices of items across the two countries; Table E.I shows that we do not find a significant difference indicating that the second term of the sampling bias is close to 1.

Table E.I: Sampling Bias Second Term: Expenditure Weights and Prices

	(1)	(2)	(3)	(4)
$\ln(\bar{p})$	-0.026	0.000	-0.026	-0.007
	(0.018)	(0.019)	(0.018)	(0.019)
$\ln(\bar{p}) \times$ Mexico	0.032	-0.019	0.032	0.023
	(0.032)	(0.012)	(0.032)	(0.029)
Observations	165	165	165	165
R-squared	0.040	0.328	0.040	0.339
Basic Heading	N	Y	N	Y
Country	N	N	Y	Y

Note: The table shows the relationship between the expenditure weights at the item level and the prices of items within a basic heading. Column (2) includes basic heading effects, Column (3) includes country effects, and Column (4) both.

To quantify the size of the third term of the sampling bias, we rely on Proposition 2. In order to compare the magnitude of $\text{cov}(\phi^{\mathbf{m}}_{\mathbf{ib}}, \ln(\bar{\mathbf{p}}^{\mathbf{m}}_{\mathbf{ib}}))$ relative to $\text{cov}(\phi^{\mathbf{u}}_{\mathbf{ib}}, \ln(\bar{\mathbf{p}}^{\mathbf{u}}_{\mathbf{ib}}))$ we estimate the following specification:

$$\phi^c_{sib} = \alpha + \beta \ln(\bar{p}^c_{sib}) \times \mathbb{1}\{c = \text{Mexico}\} + \theta^c + \lambda_i + \epsilon^c_{sib}$$

where the dependent variable are the country-specific expenditure weights at the store-item level and the independent variable are the log prices at the same level. We include country and item effects in the specification. Table E.II presents the results. It shows that the covariance of expenditure weights and prices is strongly negative for items and stores in Mexico. The results are robust after controlling for store, item, and country effects simultaneously.

Table E.II: Sampling Bias Third Term: Expenditure Weights and Prices at the Store $\times$ Item Level

	(1)	(2)	(3)
$\ln(\bar{p})$	-0.0057	-0.0055	0.0002
	(0.002)	(0.001)	(0.005)
$\ln(\bar{p}) \times$ Mexico	-0.0842	-0.0770	-0.0801
	(0.000)	(0.006)	(0.007)
Observations	419,979	419,979	416,763
R-squared	0.021	0.031	0.117
Store	N	N	Y
Item	N	Y	Y
Country	N	Y	Y

Note: The table shows the results of estimating the relationship between the country-specific expenditure weights at the store-item level and the log prices at the same level. The dependent variable is multiplied times 10^4. Column (2) includes item and country effects, Column (3) includes the same controls in addition to store effects. The standard errors are clustered at the country level.