Appendix to:
The Rise in Niche Consumption

Brent Neiman
University of Chicago

Joe Vavra
University of Chicago

November 2021

Section A of this appendix is focused on our data and empirical results while Section B elaborates on our model and theoretical results.

Appendix A. Data Appendix

We start this section of the appendix with Subsection A.1, which offers a detailed description of the Nielsen Homescan dataset, and Subsection A.2 compares the spending growth in these data to that in other datasets. Subsection A.3 then discusses the difficulty of measuring the number of aggregate varieties in these data and demonstrates the sensitivity of such measures to the treatment of products with small spending shares, while Subsection A.4 corroborates that our results on aggregate concentration are not inconsistent with concentration measures calculated using census data. Finally, we conclude with Subsection A.6, which collects a number of additional empirical results.

A.1 Detailed Data Description

Our primary data set is the AC Nielsen Homescan data, which we use to measure household-level shopping behavior.[1] As discussed in the text, our panel contains weekly household-level product spending for the period 2004-2016. The panel has large coverage, with roughly 170,000 households in over 22,000 zip codes recording prices for almost 700 million unique transactions covering a large fraction of non-service retail spending. Roughly half of expenditures are in grocery stores, a third of expenditures are in discount/warehouse club stores, and the remaining expenditures are split among smaller categories such as pet stores, liquor stores, and electronics stores.

While panelists are not paid, Nielsen provides incentives such as sweepstakes to elicit accurate reporting and reduce panel attrition. Projection weights are provided to make the sample representative of the overall U.S. population.[2] A broad set of demographic information is collected, including

[1] These data are available for academic research through a partnership with the Kilts Center at the University of Chicago, Booth School of Business. See http://research.chicagobooth.edu/nielsen for more details on the data.

[2] We use these projection weights in all reported results, but our results are similar when weighting households equally.

age, education, employment, marital status, and type of residence. Nielsen maintains a purchasing threshold that must be met over a 12-month period in order to eliminate households that report only a small fraction of their expenditures. The annual attrition rate of panelists is roughly 20 percent, and new households are regularly added to the sample to replace exiting households.

Households report detailed information about their shopping trips using a barcode scanning device provided by Nielsen. After a shopping trip, households enter information including the date and store location and scan the barcodes of all purchased items. Products are allocated by Nielsen into three levels of category aggregation: roughly 1304 "product modules", 118 "product groups", and 11 "department codes". For example, "vegetables - peas - frozen" are a typical product module within the "vegetables - frozen" product group within the "frozen foods" department, and "fabric softeners-liquid" is a typical product module within the "laundry supplies" product group within the "non-food grocery" department.

In our baseline analysis, we define a product as a UPC. UPCs are directly assigned by the manufacturer and will typically change any time there is any change in product characteristics. However, we also compute results instead defining a product as a "brand". Information on brands is constructed by Kilts/Nielsen and is more aggregated than UPCs but still very disaggregated: for example, "Pepsi" and "Caffeine Free - Pepsi" are two different brands, as are "Pepsi" and "Mountain Dew", despite the latter being produced by the same parent company. However, different flavors of Pepsi are typically all listed under the same Pepsi brand. We focus on UPCs as our baseline product definition for several reasons: 1) Most importantly, UPCs are directly assigned by the manufacturer, while the brand variable is constructed by Kilts/Nielsen. Which UPCs are grouped into more aggregate brands involves some subjective judgment, and this aggregation is not necessarily consistent across categories or time. 2) UPCs are the most fine-grained definition available and will capture relevant product changes like the introduction of new flavors which will typically not be captured with the brand-definition. 3) In order to preserve anonymity of the stores in the Nielsen sample, all generic UPCs are assigned the same brand code. This means that analysis of brand-level spending can only be done on the subset of name-brand products and must exclude the large and growing share of generic products from the sample. (see e.g. Dube *et al.* (2018)).

However, there is legitimate concern that UPCs may be too fine a notion of product when considering the concentration of household purchases, since households may view certain UPCs (for example minor differences in size or packaging for otherwise equivalent UPCs) as identical products.[3] For this reason, we show robustness to instead defining a product as a brand rather than a UPC.

Our baseline analysis focuses on annual spending and computes household market shares across products within product groups, but all results are robust to calculating household product market

[3]It is not clear that we want to classify a switch from spending $10 on Brand-X 64 oz laundry detergent and $10 Brand-X 60 oz laundry detergent to instead spending $20 on Brand-X 64 oz laundry detergent as a large increase in concentration. If UPCs become more homogeneous across time, using UPCs as our notion of product may lead to spurious changes in concentration with no substantive change in household behavior.

shares in more disaggregated product modules or more aggregated department codes. There is substantial heterogeneity across product modules in the degree of household concentration, so our analysis focuses on a set of balanced product modules. This eliminates spurious changes in concentration which might otherwise arise from changes in the set of goods sampled by Nielsen (which do not represent real changes in household's actual consumption and instead merely changes in the categories of consumption reported in Nielsen). This focus on balanced product modules reduces our sample from 118 to 107 product groups. Our analysis excludes fresh produce and other "magnet" items without barcodes since products in these categories cannot be uniquely identified and products with identical product codes in these categories can potentially differ substantially in quality. Our baseline sample includes all households and weights each household using sampling weights provided by Nielsen which are designed to make the Nielsen demographically representative of the broader U.S. population. Appendix Figure A2 shows that aggregate spending growth in our sample tracks government data on aggregate spending growth in comparable categories. Our conclusions are even stronger when instead using a balanced panel of households to eliminate household composition changes.

While our baseline sample includes all UPCs, we also show that our results hold when excluding generic/private-label products. In order to preserve anonymity of the stores in the Nielsen sample, the exact identity of generic brands in the Nielsen data is masked. There has been an increase in the private label share of all purchases over the last decade (see e.g. Dube *et al.* (2018)) so including generic spending which cannot be properly allocated to constituent UPCs might lead to spurious concentration trends. However, we show that excluding generics and calculating concentration trends for branded products produces nearly identical results.

Finally, it is also useful to discuss the potential role of online shopping for our measurement. Households in the Nielsen Homescan sample are supposed to scan barcoded purchases of purchases from online retailers in addition to the items they scan from brick-and-mortar retailers. Indeed the Nielsen panel shows a growing share of online spending across time (Figure A1). However, for the categories covered in Nielsen data, online spending is relatively unimportant, so even by the end of the sample these spending shares remain low.[4] Breaking results out further for particular categories where online spending is likely to be more and less relevant delivers no obvious interaction with concentration trends. For these reasons, we conclude that online shopping is unlikely to be of direct importance for understanding the diverging concentration trends that we document.

Figure A1: Online Spending Shares

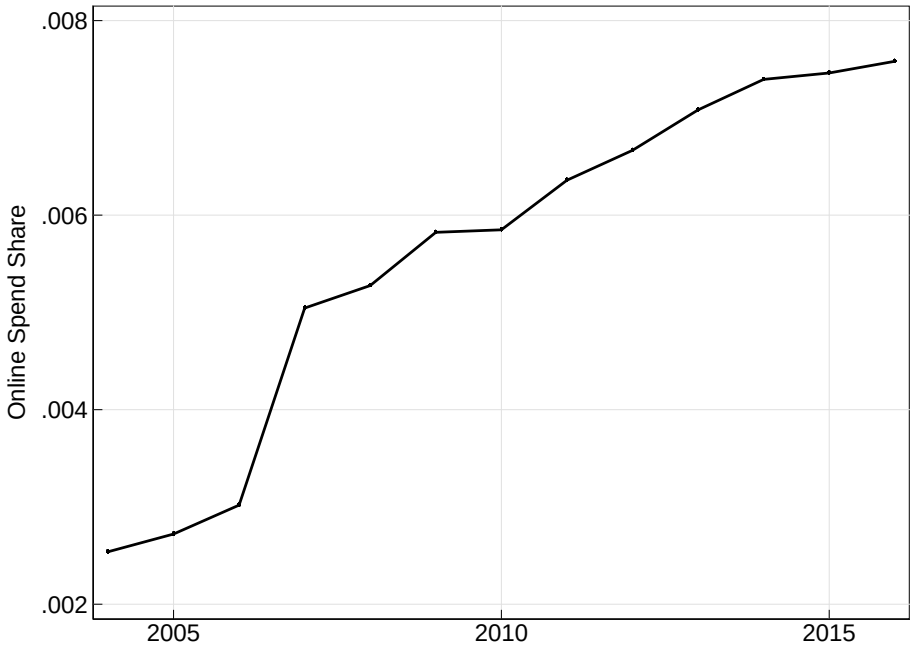
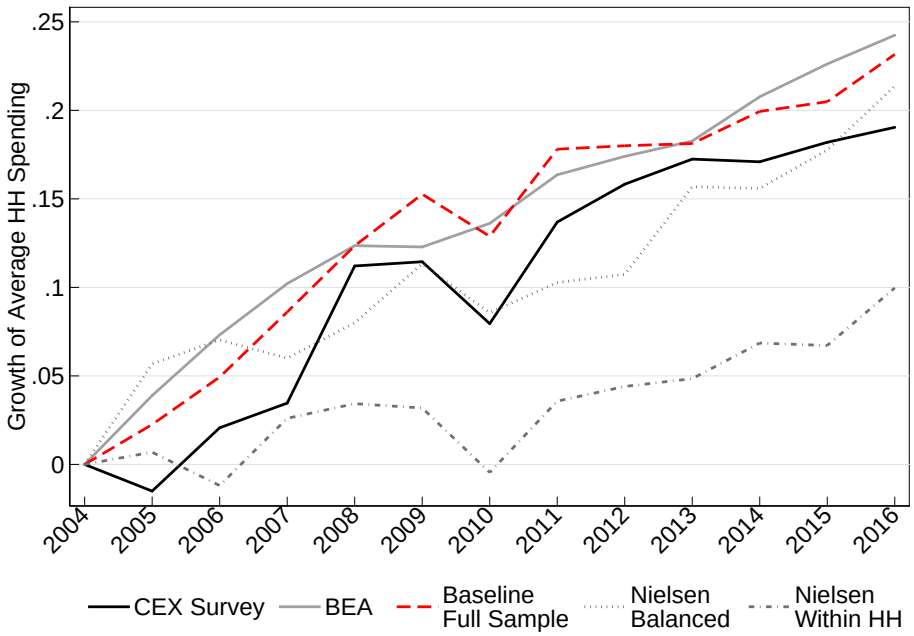


Figure A2: Household Spending in Nielsen vs. Consumer Expenditure Survey



A.2 External Spending Data

Figure A2 shows that aggregate Nielsen spending lines up well with spending growth measures from the Consumer Expenditure Survey and BEA national accounts for similar categories.[5]

However within-household spending growth is substantially less strong than overall household spending. This is likely driven by two forces: 1) The panel dimension of Nielsen is not representative of all households. The continuing households in the sample are substantially older than the overall Nielsen sample and the overall population, and we know from other research that households around retirement have declining food spending. While Nielsen provides sampling weights to make the overall sample representative of the U.S., they do not provide weights to make the panel dimension representative of the overall U.S., and the requisite demographic variables in the data to construct them ourselves do not exist. 2) There is likely attrition bias and households probably report a declining share of spending across time. This attrition bias may be particularly strong in the final year in which a household is in the sample, which could potentially explain the difference between the fully balanced and within-household spending growth patterns. If reduced reporting tends to proceed exit, then one would expect attrition bias to be less severe for households who remain in the sample for the full 12 years. Consistent with this, the balanced sample exhibits stronger spending growth than the within household sample.

For these reasons, our baseline results use the entire Nielsen homescan panel rather than focusing on a balanced panel of households. However, it is useful to compare our basic trends in the full sample to those computed using within-household variation. Figure A13 thus redoes Figure 1 using a fully balanced panel and with a specification using only the within-household changes specification.

Clearly trends are even stronger than our baseline results when using the fully balanced panel or when identifying off of within-household variation, so in this sense our baseline is conservative. We now describe several forces that might spuriously increase the within-household trend as well as some alternative forces which might spuriously flatten the full sample trend. This makes it difficult to know whether our baseline sample is likely to be understated or whether it is instead the balanced panel specification that is overstated. However, in either case, the trend is robustly positive, and our baseline sample is the one which generates more conservative results.

More specifically, the full sample trend could potentially be biased downwards because the Nielsen sampling technology changes across time, and these changes are implemented when households enter the sample. These changes in technology could obscure underlying trends in the data, but would be

[4]Online vs. brick-and-mortar spending is classified at the level of the retail chain. This means that our measure captures spending at online only retailers such as Amazon but does not classify as online spending the shopping with traditional retailers that happens to occur through their websites, such as spending at Walmart.com.

[5]It is well-known that the consumer expenditure captures a lower level of spending than the BEA and this "missing spending" has a positive trend. However this growth in missing spending mostly occurs prior to our sample period. Throughout our sample period, the CEX captures a relatively constant share of aggregate spending. This means CEX spending growth is slightly lower but broadly similar to aggregate spending growth from the BEA.

stripped out when using within-household variation. More generally, households have very different concentration levels, as shown above, so that random household entry and exit in the sample could make it more difficult to pick up underlying trends. These are both forces that might lead our baseline full sample to understate the true increase in concentration across time.

Conversely, we have shown above both that increases in spending are strongly negatively correlated with increases in concentration and that the within-household sample has spending growth much lower than in the consumer expenditure survey. To the extent that the within-household sample has spurious declining spending due to sample attrition, there is then a concern that using within household variation might lead to an upward biased trend. However, if we redo all our regression results using within household variation *controlling* for within household changes in spending, we continue to find upward trends which are stronger than in the full sample. This suggests that the stronger upward trend in the within-household results is not driven solely by the lower reported spending growth in this sample. In addition, we can also recompute results using only households in the first year in the sample. By construction, attrition bias in spending across time cannot drive any trend, since this sample has no within-household time-series variation but it still delivers an upward trend. Finally, attrition bias is less likely to be a concern for the fully balanced sample: The upward trend in the fully balanced panel is roughly linear across time, so if this upward trend was explained by attrition bias and progressive under reporting, this under reporting would need to grow at a constant rate, which seems unlikely, especially because Nielsen tries to drop households from the sample who are not reporting accurately. It seems much more likely that the biggest under reporting would occur in the first year or two in the panel as households are likely to be most enthusiastic about scanning purchases initially and then reduce scanning as it becomes more tedious across time. It would be quite surprising if enthusiasm waned at a constant linear rate across time but that households continued to participate in the homescan panel.

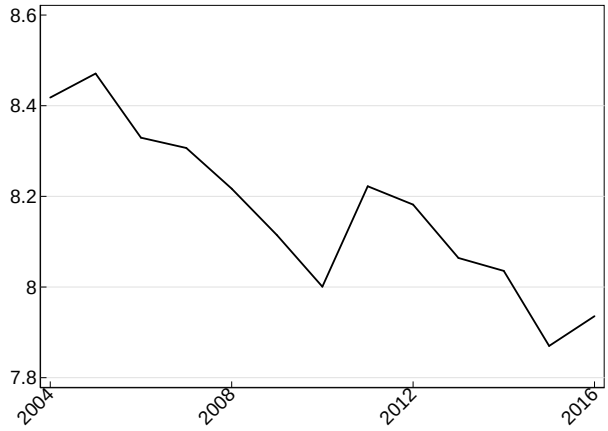
Together, we think that these results suggest the stronger upward trends using the balanced samples and the within-sample variation are not driven by spurious attrition bias. Nevertheless, we cannot fully rule out this concern. Furthermore, as discussed above the panel element of the sample is not representative since households who remain in the sample for progressive years are demographically different and not representative of the population leading total spending for this population to line up less well with aggregate spending inferred from the consumer expenditure survey. For these reasons and to be conservative, we focus on the full sample in all our baseline results but only note here that using other samples only strengthens our conclusions.

A.3 Measuring Varieties

A.3.1 Measuring Varieties Consumed Per Household

Figure 2 in the main text shows that individual households are purchasing a declining number of UPCs over time. Figure A3 shows that similar patterns hold when defining a product as a more aggregated brand rather than a UPC.

Figure A3: Ave # Brands Purchased by Each Household



It is also important to note that the typical household purchases many varieties within product categories. This motivates our modeling approach which includes love-of-variety effect at the *individual* level rather than the more standard macro model in which a representative agent has love-of-variety preferences which arise from aggregating over heterogeneous individuals who only consume a single product. We take this modeling approach because in our data context, individual households frequently purchase multiple products in narrow categories. This is true even if we focus on spending by single member households over short periods of time than our annual benchmark, so it is not driven solely by temporal aggregation or by multi-member households.

For example, focusing just on single person households, we find that for product-group weeks with positive spending, 40% have spending on 2+ UPCs, 18% have spending on 3+ UPCs, and 9% on 4+ UPCs. Again focusing on single person households but aggregating to monthly spending, we find 61% of product-group months have spending on 2+ UPCs, 37% on 3+ UPCs and 23% on 4+ UPCs.[6]

[6]These statistics weight individual households by their spending; these shares are reduced slightly if we weight households equally. We also still find frequent instances of purchasing multiple products if we define products as brands instead of UPCs but the shares are dampened by around half.

A.3.2 Measuring Varieties Consumed in the Aggregate

The variety statistics thus far focus on the number of products purchased by individual households. We now turn to a discussion of *aggregate* variety availability and show that, due to both a statistical and conceptual complication, measuring the total number of products available (or purchased) in the economy is much more challenging. Thus, we treat aggregate variety availability as unobservable in our model. Importantly, in Homescan data, we observe only the set of UPCs which are purchased by households in the panel, not the set of all products which are purchased in the economy. While the Nielsen panel is large, a large number of products nevertheless are purchased by very few households and have tiny aggregate spending. The presence of a large number of products with very small sales means that in a statistical sense, it is very hard to measure entry and exit reliably due to sampling error. If we observe a product with no sales in period t-1 and very small total spending in period t, it is difficult to tell whether the product is newly available in period t, or if we just happened to not sample a household purchasing this product in period t-1. We can show *with certainty* that the Homescan panel does not capture the full set of products available in the economy, since we can observe products which have sales in the Nielsen retail data set but no sales in Homescan. For example, of all the UPCs which are ever purchased in a Retail Panel store, 25.5% are not purchased by a single household in Homescan. One might think that we could get around this by instead measuring products in the Retail Scanner data. However, this does not solve the problem, because this data is not a census of all stores. For example, we can see that 22.2% of UPCs which are purchased in Homescan are not sold in any store in the retail panel.

The more conceptual challenge, which would not be solved even if we had a full census of all U.S. product sales, is that our model implies a distinction between products which are available and products which are purchased. We interpret products which are available but have no sales in our model as failed products. However, this is clearly an abstraction, and even the worst failed products will likely have tiny, but not actually zero sales. This means that even if there were no sampling issues related to products with small spending, we might still want to include some minimum aggregate spending threshold in order to "count" a product in the data.

Table A1 shows that the treatment of products with tiny spending in the Nielsen data indeed makes a huge difference for measures of aggregate varieties (both in levels and in growth rates), which is why we choose to treat this as an unobservable object in our model. For example, this table shows that although they represent only 2% of total spending, roughly half of all UPCs in the Homescan data have total annual spending across all households of less than $25. Excluding products with small aggregate spending also leads to large changes in measured variety growth: counting products with even extremely tiny aggregate spending, delivers growth of 6.2% from 2004 to 2016, while dropping products with very tiny spending (which again are more sensitive to sampling error and interpretation

issues) raises measured growth to 20-30%.[7] Thus, while the data paints a robust pattern that the number of products is large and growing, exact product counts and growth rates are too uncertain to be usable as inputs to our model or resulting welfare inference.

Table A1: Effect of Products with Small Aggregate Spending on Statistics

Agg Spend Threshold	Share Spend> Threshold	UPCs per category 2004	UPCs per category 2016	UPCs per category % change	$\mathcal{H}^{HH}_{2004}$	$\mathcal{H}^{HH}_{2016}$	$\mathcal{H}^{HH}$ % change	$\mathcal{H}^{Agg}_{2004}$	$\mathcal{H}^{Agg}_{2016}$	$\mathcal{H}^{Agg}$ % change	Ω_{2004}	Ω_{2016}	Ω % change
\$0	100%	9248	9820	6%	0.262	0.284	8%	0.0036	0.0028	-21%	16.6	15.6	-6%
\$25	98%	4362	5193	19%	0.268	0.290	8%	0.0038	0.0030	-21%	15.8	15.0	-5%
\$50	95%	3153	3856	22%	0.275	0.296	7%	0.0039	0.0031	-22%	15.3	14.6	-5%
\$250	83%	1206	1555	29%	0.304	0.324	6%	0.0051	0.0039	-22%	13.2	12.7	-4%

In contrast, the statistics which are the focus of our analysis (household and aggregate Herfindahls as well as the average number of products purchased by individual households) are very robust to the treatment of these products with small aggregate spending, since these statistics depend much more on products with substantial spending. Growth rates of these variables (which are more important for our model inference) are even more stable across these spending thresholds, changing by at most a couple percentage points when moving from no aggregate spending threshold to a fairly restrictive threshold. An advantage of our modeling framework is thus that we can infer product availability changes and their welfare consequences using these observable statistics even though we cannot reliably measure product availability itself.

Most importantly Table A1 shows that our model inference for variety availability and welfare are almost completely unaffected by the behavior of these products with small aggregate spending. Performing inference on statistics constructed using spending on all products produces nearly identical

[7] Note that for all of the calculations in Table A1, we compute statistics using a random subset of the Homescan Panel with a constant number of households per year so that statistics are not affected by changes in the panel size.

Table A2: Effect of Products with Small Aggregate Spending on Model Implied Annual Growth Rates

Agg Spend Threshold	j^* growth	$\tilde{N}$ growth	η growth	Utility growth from N	Utility growth from N, F, σ
\$0	2.0%	4.5%	-0.18%	0.56%	0.45%
\$25	2.0%	4.6%	-0.21%	0.57%	0.46%
\$50	2.1%	4.6%	-0.21%	0.57%	0.47%
\$250	2.1%	4.5%	-0.26%	0.56%	0.47%

conclusions to inference performed on statistics which exclude products with tiny or modest aggregate spending. For example, in all cases, the model implies that the annual growth in consumed varieties (j^*) is always 2-2.1%, and that annual welfare growth when fully accounting for all of the time-trends in the data is 0.45%-0.47%. Thus, none of our model conclusions are affected by the behavior of the large set of UPCs in Nielsen with negligible spending.

A.4 Census Concentration of Production

A large and growing literature uses production data from the Census to show that the concentration of production has been broadly increasing from 1982-2012. For example, Autor *et al.* (2017) calculates industry concentration within 4-digit industries, and averages this up to 6 major sectors and shows that various concentration measures have all increased when comparing 1982 to 2012. In this section we explore the relationship between the concentration measures in our paper and this large literature and argue that relevant comparisons from Nielsen data are broadly consistent with this Census based literature.

First, it is important to note that the concentration notions we emphasize in our paper are conceptually distinct along a number of important dimensions from the concentration of firms or establishments studied using census data. Most importantly, we are measuring the concentration of spending over very detailed UPCs (or slightly coarser but still highly disaggregated brands). This is a fundamentally much more disaggregated notion of concentration than that studied with production data, since firms can potentially produce tens, hundreds or even thousands of different products. For example, in our data Procter and Gamble produces over 40,000 unique UPCs, L'Oreal produces over 28,000 UPCs and General Mills, Unilever, and Kraft Heinz all produce 10,000-20,000 UPCs.[8]

Furthermore, the categories within which we calculate concentration are also more disaggregated than those in typical Census-based calculations and also cover a more narrow subset of production. For example, the broad manufacturing sector in Autor *et al.* (2017) covers 86 4-digit industries within which concentration is computed. However, of these 86 industries only a small subset produce in categories which are covered by Nielsen (for example NAICS Code 3111 "Animal Food Manufacturing") while most are in production industries which have no overlap with Nielsen categories (for example NAICS Code 3336 "Engine, turbine, and power transmission equipment manufacturing" or NAICS Code 3365 "Railroad rolling stock manufacturing").

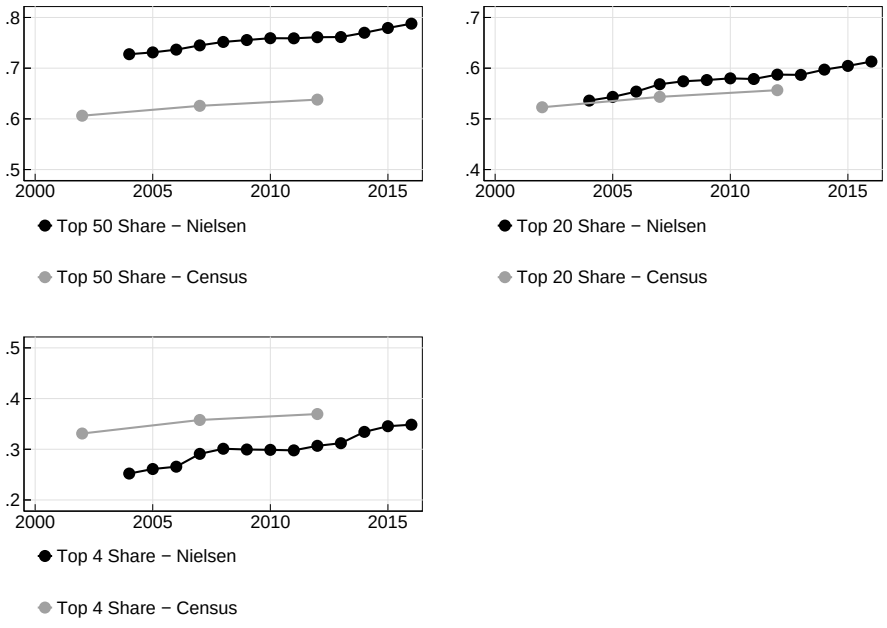
Finally, it is important to note that our sample covers the period 2004-2016 while census data starts in 1982 and is last available in 2012. The exact timing of concentration trends in Census data varies substantially, with many sectors exhibiting increases primarily in the period prior to our sample period.

[8]It is also worth noting that our "household" concentration measures have no analogue in the Census literature even if we were measuring producer rather than product concentration.

Since they are conceptually different notions, this means the aggregate product concentration trends which we emphasize in the body of the paper should not be directly compared to production concentration trends in Census. However, we can construct concentration measures using the Nielsen data which *are* more comparable with Census calculations and that can be used to explore the external validity of our data. We now explore these comparisons.

Since households in the Nielsen sample report the retail chain in which they shop, we can aggregate up total spending to compute a Nielsen based measure of spending at each retail chain and resulting retailer concentration. This can then be compared to the concentration of retail trade in Census data. Specifically, since the Nielsen sample is focused on grocery and drug store spending, in the Census we use firm concentration numbers only from NAICS Code 445 "Food and beverage stores" and 446 "Health and personal care stores" and weight the publicly available Census concentration numbers for these two sectors using their relative share of sales. This clearly does not provide a precise match between the retail establishments covered in Nielsen and Census so we should not expect numbers to line up exactly, but Figure A4 shows that that Nielsen data broadly matches the level of retail spending accounted for by the Top 4, Top 20 and Top 50 firms as well as the upward trend in retail concentration.

Figure A4: Retail Trade Concentration



We can also perform a similar exercise by allocating UPC-level spending up to the manufacturer. When manufacturers produce a new product, it is assigned a barcode by the company GS1, which then maintains a database which can be used to link UPCs to manufacturers. This lets us aggregate product spending up to a measure of manufacturer spending, with two important caveats:

First, the link from UPCs to parent companies is sometimes inconsistent. For example, Gillette and Old Spice were both acquired in the past by Proctor and Gamble, and the UPCs for Gillette and Old Spice products both map to Proctor and Gamble. However, Ben and Jerry's was acquired by Unilever in 2000, yet UPCs for these products are assigned to the "Ben and Jerry's Homemade Inc" firm name rather than to the Unilever parent company. Similarly, Goose Island Beer UPCs are assigned to "Goose Island Beer Company" even though this firm was acquired by InBev in 2011. To the extent that some UPCs are assigned to subsidiaries rather than parent companies, our Nielsen based measure of manufacturer concentration will be biased downwards.

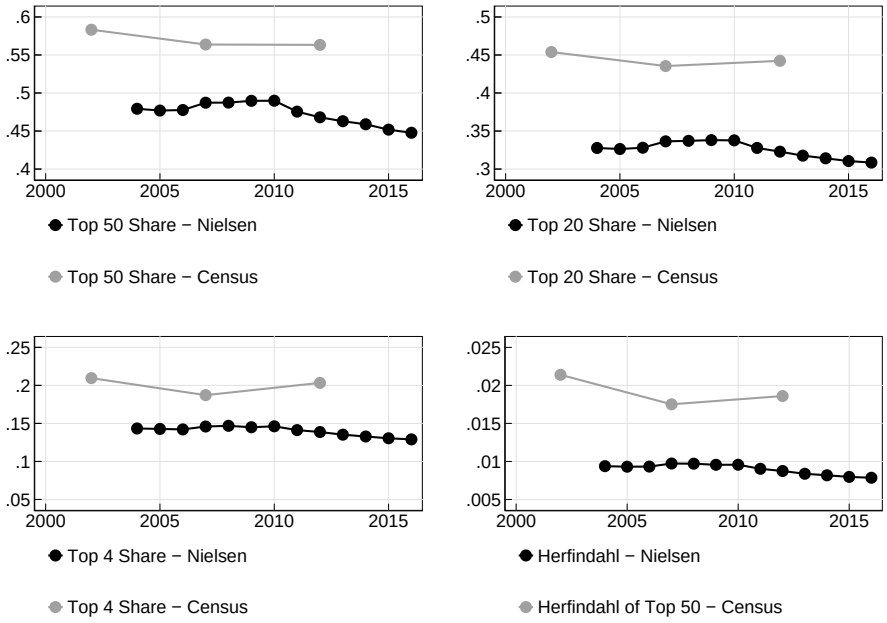
Second, UPCs for store-brand products map to the retailer rather than the actual manufacturer of the product. For example, Costco's "Kirkland" store-brand barcodes all map to "Costco", even though Costco does not actually produce most of these products. Although sometimes the actual producer can be identified (for example Kirkland Coffees are advertised as being roasted by Starbucks), this information is typically a trade-secret.[9] This means that we cannot measure the producer for most generic products, and as a result we must drop these products when aggregating up UPCs to manufacturers and focus only on branded products. To the extent that the production of generic products is proportional to the production of branded products, this will have no effect on concentration. However, it is likely that generic products are disproportionately produced by larger manufacturers, so dropping generic products is likely a second force that will bias our Nielsen based measures of manufacturer concentration downwards.

To again focus the comparisons on the most relevant producers, we keep NAICS codes 311 and 312 "Food Manufacturing" and "Beverage and Tobacco Product Manufacturing" from the Census data and weight these concentration measures by their relative sales shares. Figure A5 shows that despite the above concerns, Nielsen data again broadly matches Census data, producing similar levels of manufacturer concentration and a flat to mild downward trend.

Overall the results in these two subsections give us confidence that the Nielsen data is largely in line with external evidence on aggregate spending and with Census data on producer concentration.

[9]Note that this specific anecdotal example is not drawn from the actual Nielsen micro data. License agreements prevent any disclosure of information about specific retailers or individual stores from this data.

Figure A5: Manufacturer Concentration



A.5 Additional Corroboration of Model Fit at Household Level

Figure A6 shows a number of additional rank comparisons of our main comparison of the household model fit discussed in Figure 7.

An alternative to the test of household-level model fit that we present in Figure 7 in the main text is to use the testable prediction from equation (18) that $\mathcal{H}_{i,c}^{HH}$ is proportional to $1/|\Omega_{i,c}|$. Indeed, when we pool categories, years, and households and regress $\ln|\Omega_{i,c}|$ on $-\ln\mathcal{H}_{i,c}^{\text{HH}}$ (with category-year fixed effects), we get a coefficient of 0.89, which is close to the model-consistent value of 1, and a large R^2 of 0.82. The upper left panel of Figure A7 shows a binscatter (with category-year fixed effects) of the 54 million observations underlying this regression to demonstrate that linearity with a coefficient of 1 is a close approximation to the raw data.[10] In the upper right panel, we estimate these regressions separately for each category in 2016 and plot a histogram of the estimated slopes. The values are largely clustered around the model-consistent value of 1.

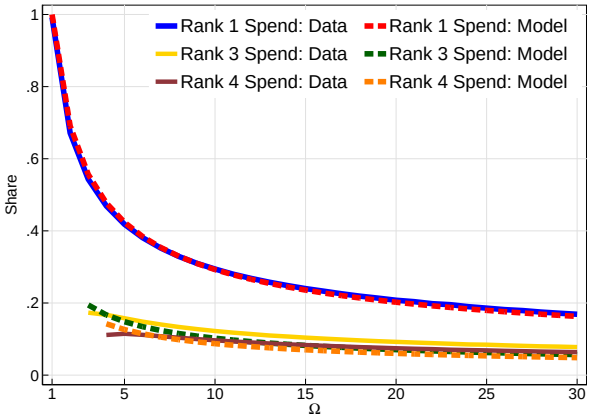
Next, rather than estimating the slope, we constrain it to equal 1 and back out the η_c values implied for each category. The model imposes the restriction that $0 < \eta_c < 1$ and the bottom left panel of Figure A7 shows that this restriction is satisfied in every category. The values of η_c range from lows of 0.08 (Baby Food) and 0.10 (Carbonated Beverages) to highs of 0.69 (Greeting Cards) and 0.97 (Yeast).[11]

[10] This specification has large explanatory power even though it only allows η to vary across category-years and not across households. With arbitrary heterogeneity in η across households within category-years, there would be as many parameters as observations so it would be trivial to perfectly fit the data.

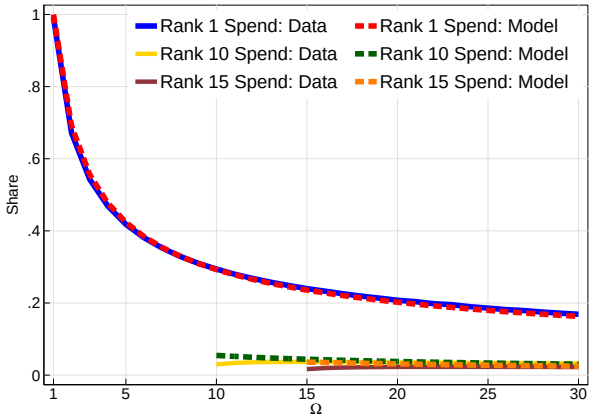
[11] The value of 0.08 for baby food implies that the typical household in this category has spending which is almost 4 times

Figure A6: Alternative Concentration Measures

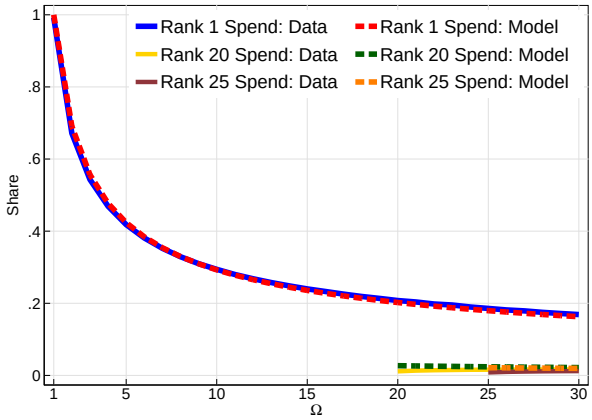
(a): Model vs Data: Additional Ranks 3 and 4



(b): Model vs Data: Additional Ranks 10 and 15



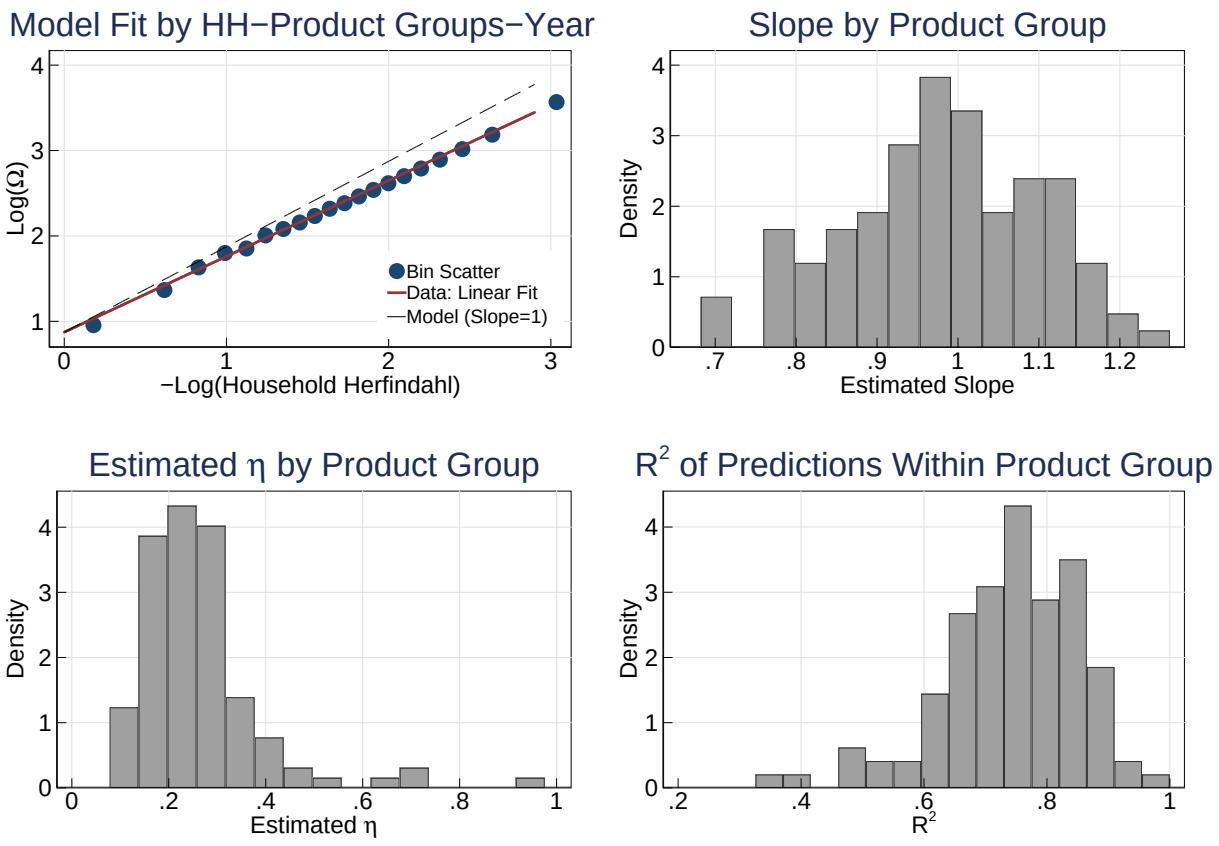
(c): Model vs Data: Additional Ranks 20 and 25



Finally the lower right panel shows that the R^2's from these restricted regressions are generally high. Overall, we conclude that the empirical relationship between household-level concentration measures and the number of consumed products is consistent with the relationships implied in our model.

more concentrated than if that household spent evenly on all the baby food products they consumed, while the value of 0.97 for yeast implies that household spending in that category is essentially evenly divided across products. With homogeneous tastes across products (i.e. $\theta \to \infty$) – the setup in many standard models – we cannot capture this large extent of sectoral heterogeneity in concentration.

Figure A7: Model Fit on Household-Category Data



A.6 Additional Empirical Results

Figure A8: Concentration Trends: Excluding Generics

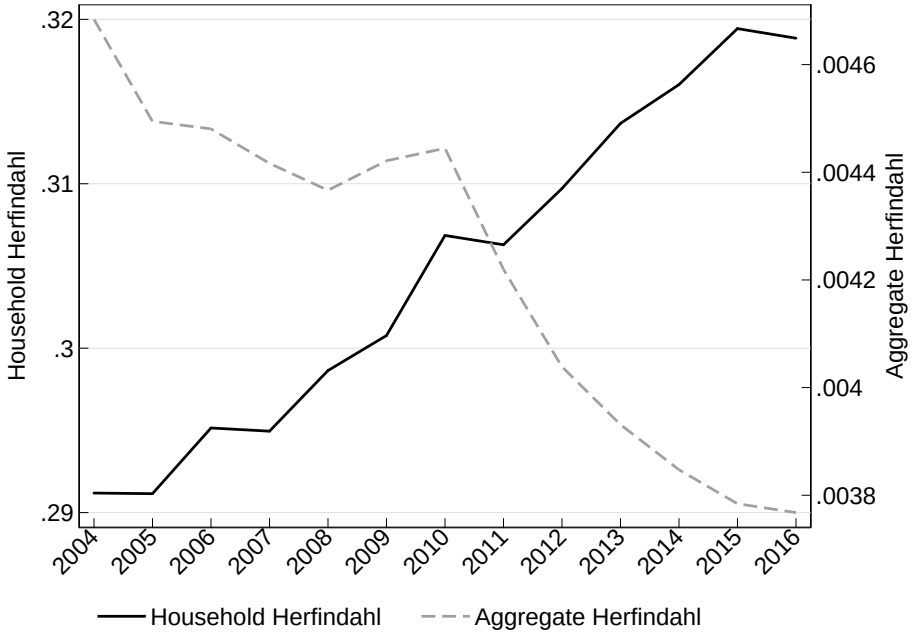


Figure A9: Concentration Trends: Including Category Composition Changes

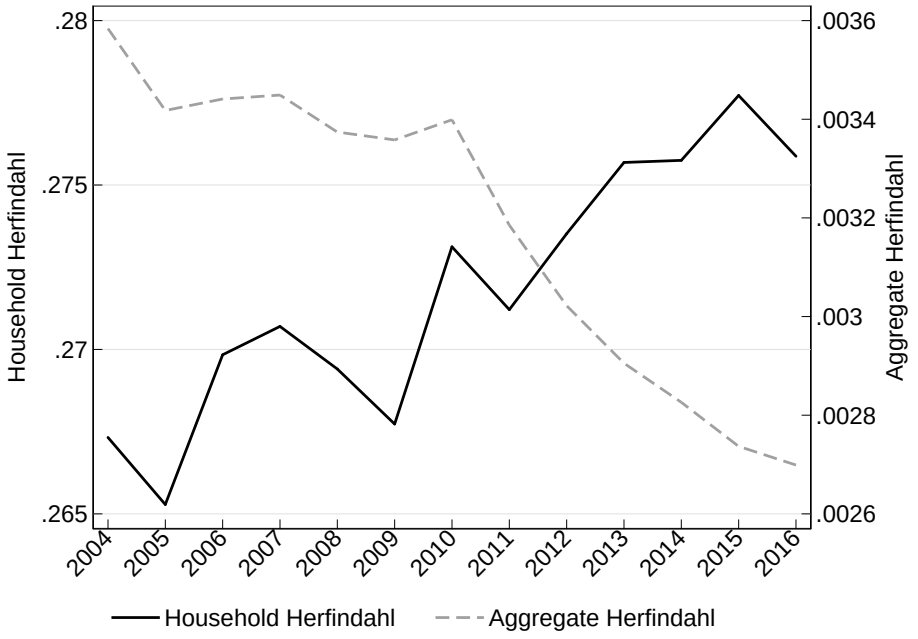


Figure A10: Concentration Trends: Brand Instead of UPC

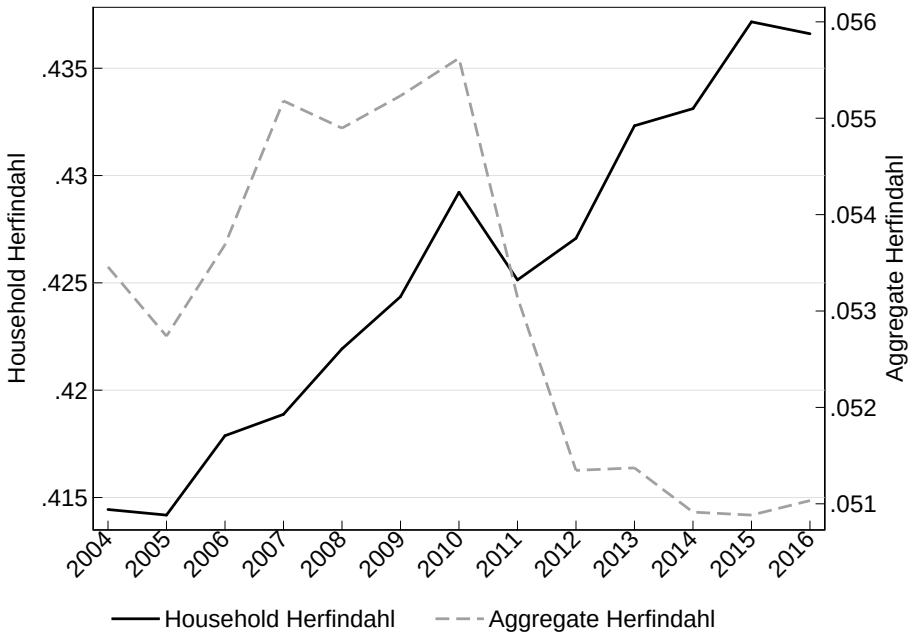


Figure A11: Concentration Trends: Product Module instead of Group

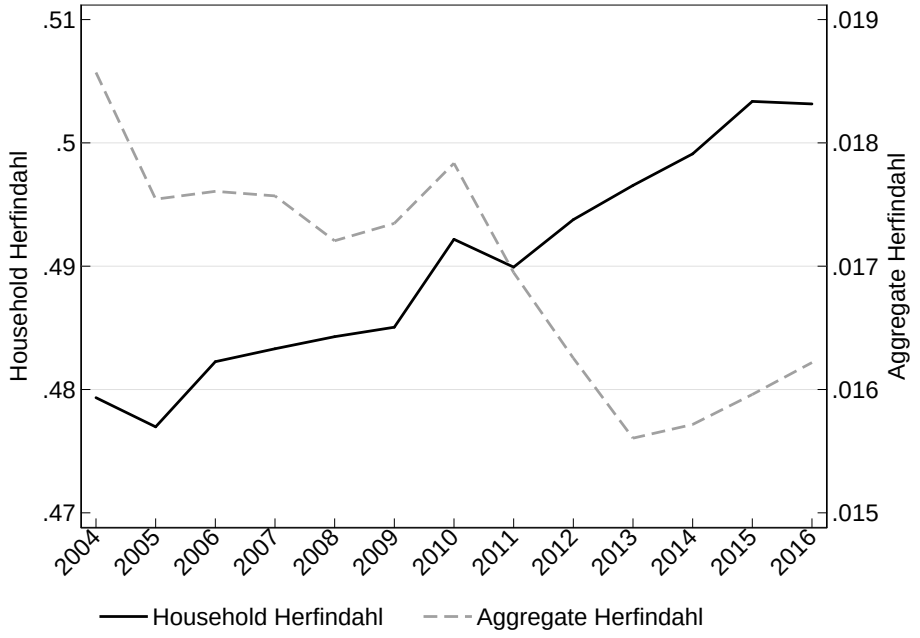
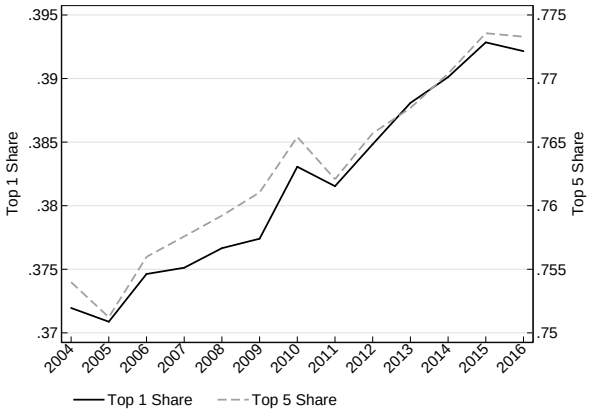


Figure A12: Alternative Concentration Measures

(a): Household Shares on Top Products



(b): Aggregate Shares on Top Products

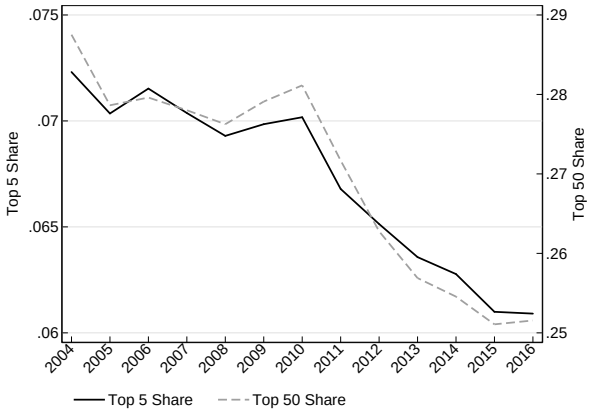
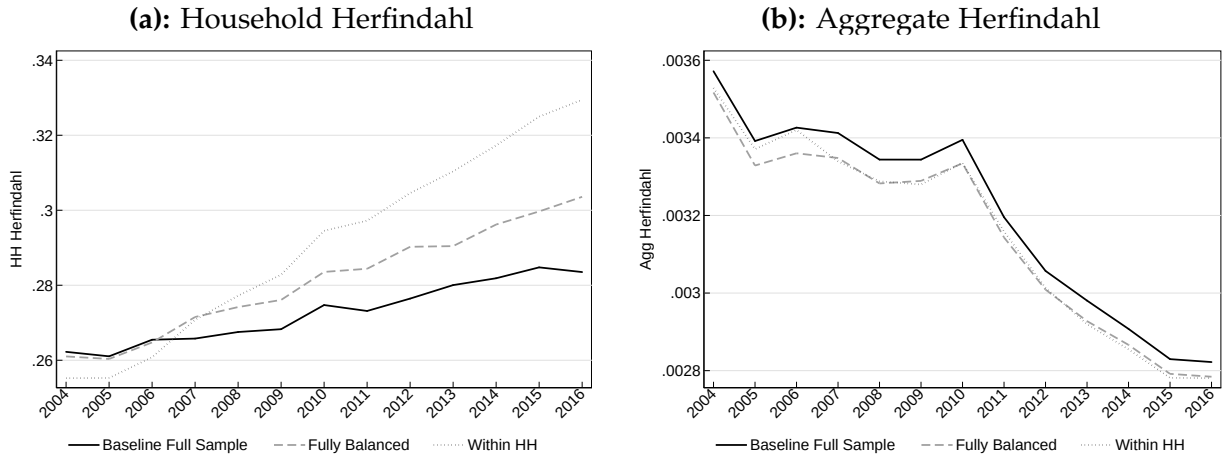


Figure A13: Concentration Trends for Different Samples



This figure recomputes Figure 1 for the baseline sample, a fully balanced sample and using only within household variation. The within-household variation specification calculates changes in concentration within household between years, averages these across household-categories and cumulates these changes over time beginning from the baseline level. The baseline sample averages individual *levels* instead of individual *changes* in household concentration. This means that it is more representative since it includes households only in the Nielsen panel for a single year, but it can change as households enter and leave the panel. Figure A2 shows that the baseline sample is a better fit to aggregate spending trends.

Figure A14: 2004-2016 Concentration Growth by Household Size

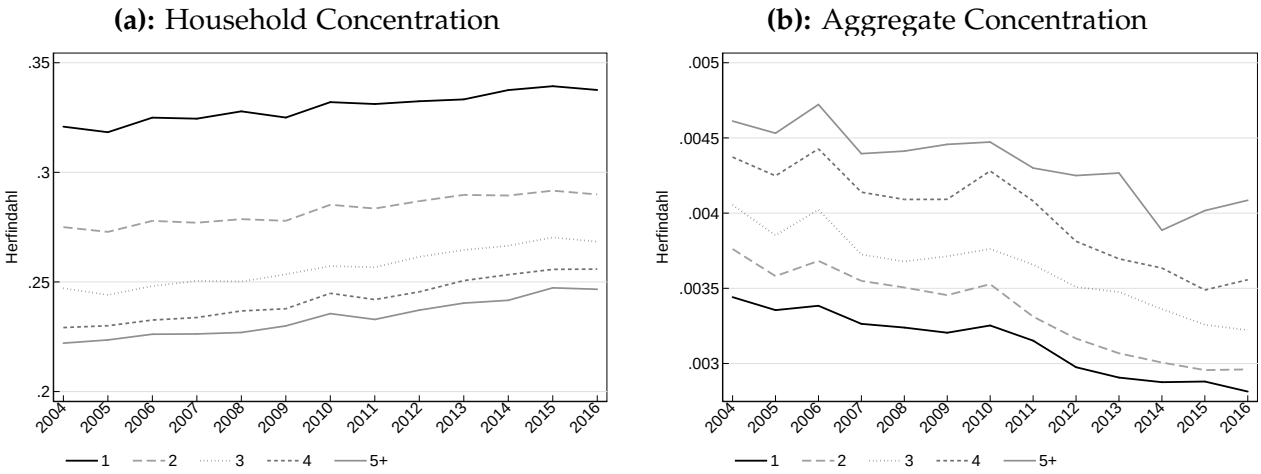
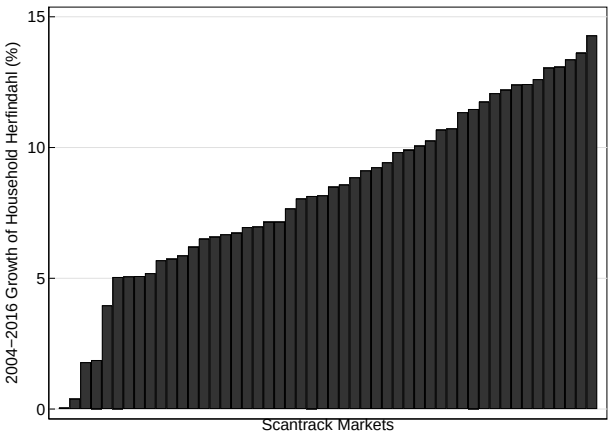


Figure A15: 2004-2016 Concentration Growth Within Location

(a): Household Concentration



(b): Aggregate Concentration

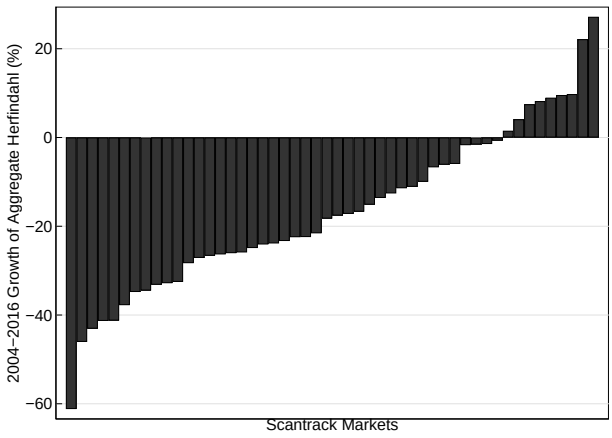
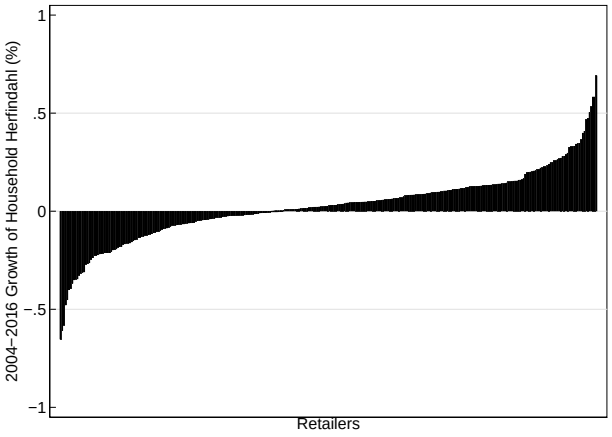


Figure A16: 2004-2016 Concentration Growth Within Retailer

(a): Household Concentration



(b): Aggregate Concentration

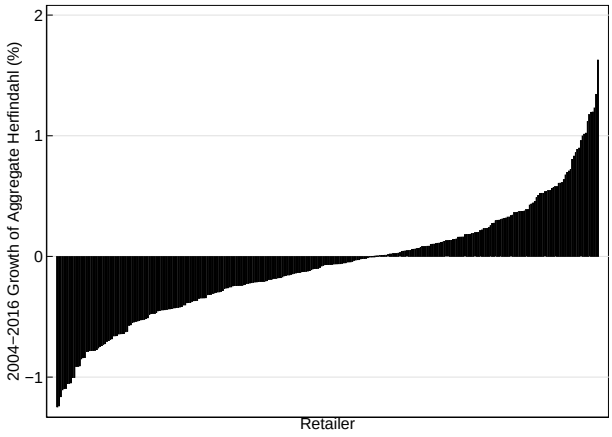
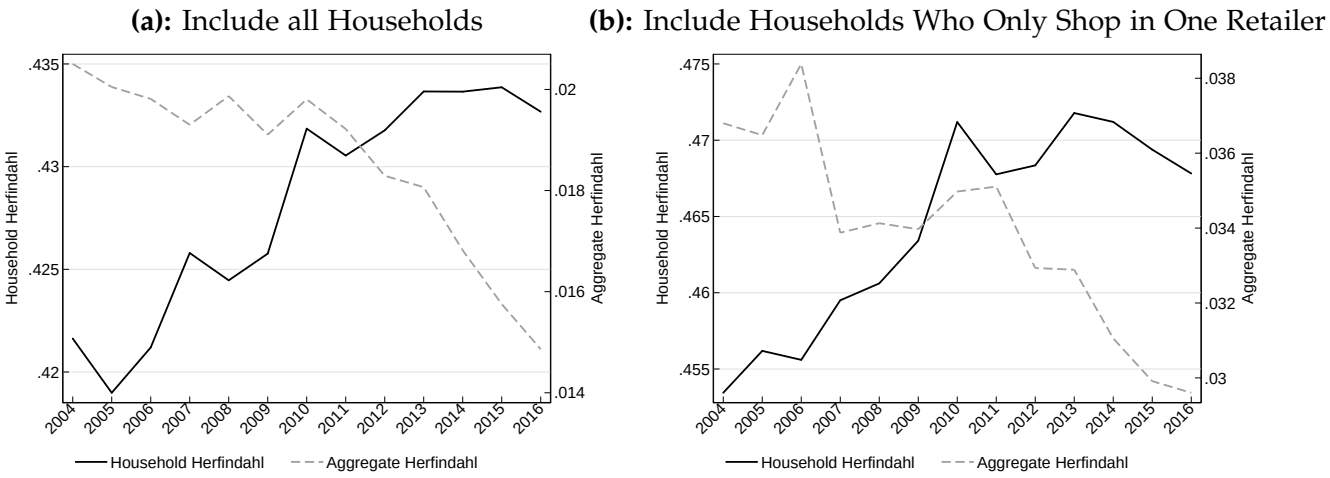
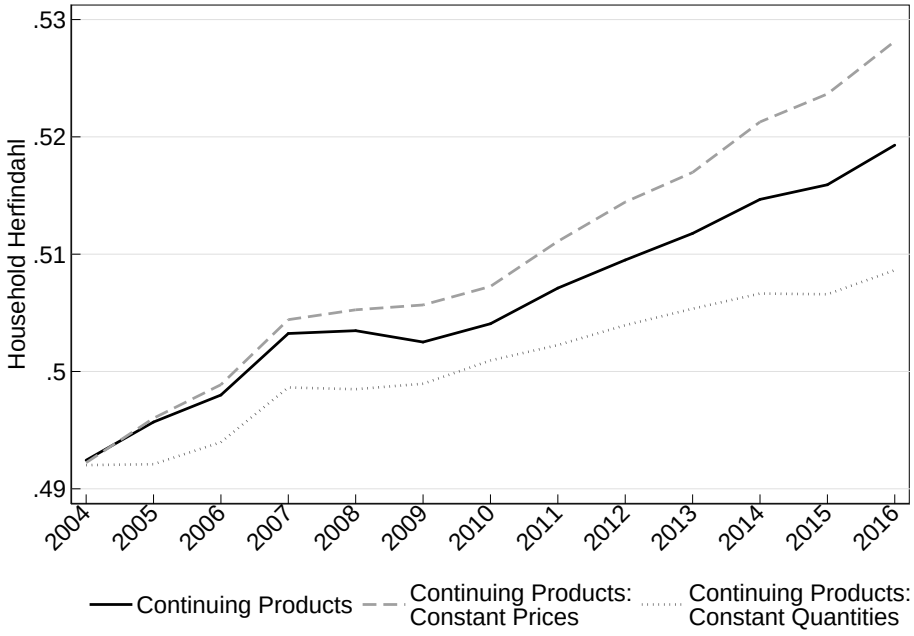


Figure A17: Within-Retailer Concentration Trends



This figure recomputes Figure 1 but defining market shares within retailer-categories instead of within categories. That is, we calculate household and aggregate concentration within individual retailer-category pairs and then average across all retailers and categories. The left panel includes all households while the right uses category spending only of households who shop in a single retailer for a given category to eliminate any composition effects from shifting expenditure across retailers within categories.

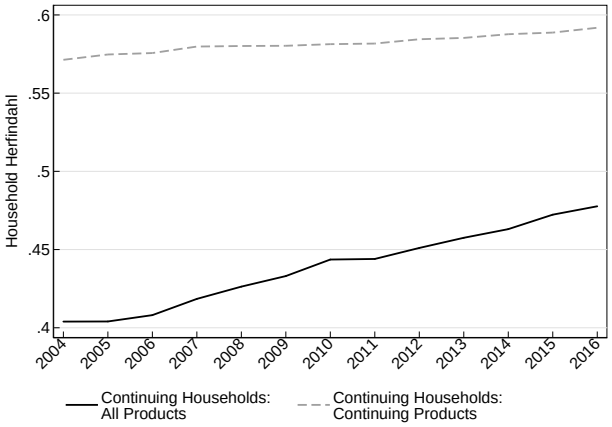
Figure A18: Intensive Margin P v. Q effects for UPCs



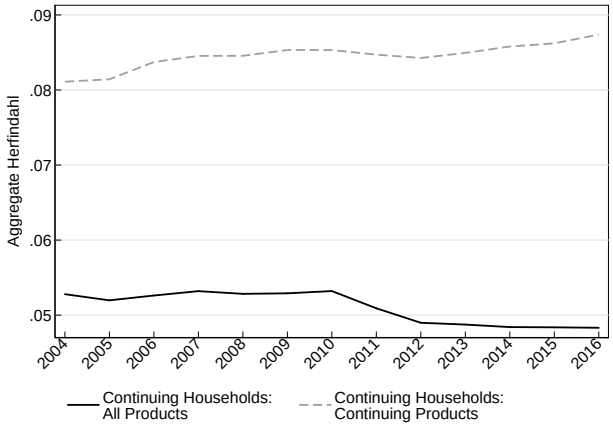
This figure recomputes the continuing product line in figure 6 but now either holding prices or holding quantities constant between t and $t+1$ for each household.

Figure A19: 2004-2016 Concentration growth for continuing vs. all products (brands)

(a): Household Herfindahl



(b): Aggregate Herfindahl



Appendix B. Theory Appendix

We start this section of the appendix with Subsection B.1, which provides detailed derivations of the main expressions presented in the body of the paper. Subsection B.2 offers additional results related to the simulation of our model. Finally, Subsection B.3 presents a version of our model where demand is assumed to be linear, rather than exhibiting a constant elasticity of substitution.

B.1 Derivation of Key Model Equations

In this appendix subsection, we provide detailed derivations of our expression for the price index P, the Household Herfindahl $\mathcal{H}^{\text{HH}}$, the CDF characterizing rank-values $R(r)$, the aggregate market share of a product s_j, the Aggregate Herfindahl $\mathcal{H}^{\text{Agg}}$, and elasticities of demand ε_j and markups at the product μ_j and aggregate levels μ^{Agg}.

B.1.1 Deriving the Price Index in Equation (14)

We start by showing how we can write $U_i = E/P - F\left(|\Omega_i|\right)^{\epsilon}$, where P is defined as in equation (14). The Lagrangian for the household consumption problem can be written as:

$$\mathcal{L} = \left(\int_{k\in\Omega_i} \left(\gamma_{i,k} C_{i,k}\right)^{\frac{\sigma-1}{\sigma}} dk\right)^{\frac{\sigma}{\sigma-1}} - F \times \left(|\Omega_i|\right)^{\epsilon} - \lambda\left(\int_{k\in\Omega_i} p_k C_{i,k} dk - E\right),$$

where λ is the multiplier on the budget constraint. Taking $|\Omega_i|$ as given and differentiating with respect to $C_{i,k}$ and setting equal to zero, we get:

$$\gamma_{i,k} C_{i,k} = \left(\lambda \frac{p_k}{\gamma_{i,k}} \frac{\sigma}{\sigma-1}\right)^{-\sigma},$$

for each good $k \in \Omega_i$. Taking the ratio of this first-order condition for good k to some other good j, rearranging, and integrating over goods k, we get:

$$E = \int_{k\in\Omega_i} C_{i,k} p_k dk = C_{i,j} \left(p_j\right)^{\sigma} \left(\gamma_{i,j}\right)^{1-\sigma} \left(\int_{k\in\Omega_i} \left(p_k\right)^{(1-\sigma)} \left(\gamma_{i,k}\right)^{\sigma-1} dk\right).$$

Defining $P_i = \left(\int_{k\in\Omega_i} \left(\tilde{\gamma}_{i,k}\right)^{\sigma-1} dk\right)^{\frac{1}{1-\sigma}}$, we can then re-write this expression as:

$$\gamma_{i,j} C_{i,j} = E \left(P_i\right)^{\sigma-1} \left(\tilde{\gamma}_{i,j}\right)^{\sigma}.$$

Next, we can raise both sides to the power $(\sigma-1)/\sigma$ and integrate over goods j to show that:

$$\int_{j\in\Omega_i} \left(\gamma_{i,j} C_{i,j}\right)^{\frac{\sigma-1}{\sigma}} dj = \left(E\right)^{\frac{\sigma-1}{\sigma}} \left(P_i\right)^{\frac{(\sigma-1)^2}{\sigma}} \int_{j\in\Omega_i} \left(\tilde{\gamma}_{i,j}\right)^{\sigma-1} dj = \left(E\right)^{\frac{\sigma-1}{\sigma}} \left(P_i\right)^{\frac{(\sigma-1)^2}{\sigma}} \left(P_i\right)^{1-\sigma}.$$

Finally, raising both sides to the $\sigma/(\sigma-1)$, gives:

$$\left(\int_{k\in\Omega_i}\left(\gamma_{i,k}C_{i,k}\right)^{\frac{\sigma-1}{\sigma}}dk\right)^{\frac{\sigma}{\sigma-1}}=E/P_i,$$

which will hold for any distribution of $\tilde{\gamma}_{i,k}$, regardless of the combinations of p_k and $\gamma_{i,k}$ that give rise to it.

Next, note that under the assumption that $G(y)$ is a Pareto distribution, the cutoff good is characterized by:

$$\frac{|\Omega_i|}{N}=\left(\frac{b}{\tilde{\gamma}_i^*}\right)^{\theta},$$

which implies:

$$\tilde{\gamma}_i^*=N^{\frac{1}{\theta}}\left(|\Omega_i|\right)^{-\frac{1}{\theta}}b. \tag{A1}$$

From the definition of P_i, we therefore have:

$$\begin{aligned}P=P_i &= \left(\int_{k\in\Omega_i}\left(\tilde{\gamma}_{i,k}\right)^{\sigma-1}dk\right)^{\frac{1}{1-\sigma}}=\left(N\int_{\tilde{\gamma}_i^*}^{\infty}y^{\sigma-1}dG(y)\right)^{\frac{1}{1-\sigma}}\\ &= N^{\frac{1}{1-\sigma}}\left(\frac{1}{\sigma-1-\theta}y^{\sigma-1-\theta}\Big|_{\tilde{\gamma}_i^*}^{\infty}\right)^{\frac{1}{1-\sigma}}\theta^{\frac{1}{1-\sigma}}b^{\frac{\theta}{1-\sigma}}=\left(1+\frac{1-\sigma}{\theta}\right)^{\frac{1}{\sigma-1}}b^{\frac{\theta}{1-\sigma}}N^{\frac{1}{1-\sigma}}\left(\tilde{\gamma}_i^*\right)^{\frac{\theta}{\sigma-1}-1}.\end{aligned}$$

Substituting the value for $\tilde{\gamma}_i^*$ from equation (A1), we get:

$$P=P_i=\underbrace{\left(1+\frac{1-\sigma}{\theta}\right)^{\frac{1}{\sigma-1}}b^{-1}}_{\text{Average Price}}\times\underbrace{\left(|\Omega_i|\right)^{\frac{1}{1-\sigma}}}_{\text{Variety Effects}}\times\underbrace{\left(\frac{|\Omega_i|}{N}\right)^{\frac{1}{\theta}}}_{\text{Selection Effects}}.$$

B.1.2 Deriving the Household Herfindahl $\mathcal{H}^{\text{HH}}$ in Equation (18)

We have:

$$\mathcal{H}^{\text{HH}}=\mathcal{H}_i^{\text{HH}}=N\int_{\tilde{\gamma}_i^*}^{\infty}\left(P_iy\right)^{2(\sigma-1)}G(y)dy=P^{2(\sigma-1)}\frac{N\theta b^{\theta}}{\theta-2\left(\sigma-1\right)}\left(\tilde{\gamma}^*\right)^{2(\sigma-1)-\theta}.$$

Substituting in the definition of $\tilde{\gamma}_i^*$ from equation (A1), we have:

$$\mathcal{H}^{\text{HH}}=P^{2(\sigma-1)}\frac{\theta}{\theta-2\left(\sigma-1\right)}N^{\frac{2(\sigma-1)}{\theta}}|\Omega|^{1-\frac{2(\sigma-1)}{\theta}}b^{2(\sigma-1)},$$

and substituting in the definition of P from equation (14), we get:

$$\mathcal{H}^{\text{HH}}=\left(1+\frac{1-\sigma}{\theta}\right)^{2}\frac{1}{1-\frac{2(\sigma-1)}{\theta}}\frac{1}{|\Omega|}.$$

Defining $\eta = 1 - \frac{2(\sigma-1)}{\theta}$, we then have:

$$\begin{aligned}\mathcal{H}^{\text{HH}} &= \frac{1}{\eta}\left(1+\frac{1-\sigma}{\theta}\right)^2\frac{1}{|\Omega|} = \frac{1}{\eta}\left(1-\frac{2(\sigma-1)}{\theta}+\frac{\sigma-1}{\theta}\right)^2\frac{1}{|\Omega|} \\ &= \frac{1}{\eta}\left(\eta+\frac{\sigma-1}{\theta}\right)^2\frac{1}{|\Omega|} = \frac{(\eta+1)^2}{4\eta}\frac{1}{|\Omega|}.\end{aligned}$$

B.1.3 Deriving the CDF $R(r)$ in Equation (21)

The goal here is to find the CDF of the rank value $r_{i,j} = (1-\alpha)\,j + \alpha x_{i,j}$ when these values are pooled across households i and products j. As a first step, let's pool only across households and solve for the conditional CDF for each product $j \in (0, N]$:

$$R_j(r) = Pr\left[(1-\alpha)\,j + \alpha x_{i,j} \leq r\right] = Pr\left[x_{i,j} < \frac{r-(1-\alpha)\,j}{\alpha}\right].$$

This yields:

$$R_j(r) = \begin{cases} 0, & 0 \leq r < (1-\alpha)\,j \\ \dfrac{r-(1-\alpha)\,j}{\alpha N}, & (1-\alpha)\,j \leq r < (1-\alpha)\,j + \alpha N \\ 1, & (1-\alpha)\,j + \alpha N \leq r \leq N. \end{cases}$$

We then can get the unconditional CDF by integrating these conditional CDFs across all products: $R(r) = \frac{1}{N}\int_0^N R_j(r)dj$. As an intermediate step, it will be useful to assume that $\alpha < 0.5$, which implies that $\alpha < 1-\alpha$, and to re-write the boundaries on the parameter space that define the three regions of the conditional CDF as follows:

$$R_j(r) = \begin{cases} 0, & min\left(\dfrac{r}{1-\alpha}, N\right) \leq j < N \\ \dfrac{r-(1-\alpha)\,j}{\alpha N}, & max\left(0, \dfrac{r-\alpha N}{1-\alpha}\right) \leq j < min\left(N, \dfrac{r}{1-\alpha}\right) \\ 1, & 0 \leq j \leq max\left(0, \dfrac{r-\alpha N}{1-\alpha}\right), \end{cases}$$

where the *min* and *max* conditions come from the restriction that $j \in (0, N]$.

We can then start by calculating the CDF in the first region – where $0 \leq r < \alpha N$ – as:

$$\begin{aligned}R(r) &= \frac{1}{N}\int_0^N R_j(r)dj = \frac{1}{N}\int_0^0 1\times dj + \frac{1}{N}\int_0^{\frac{r}{1-\alpha}}\frac{r-(1-\alpha)\,j}{\alpha N}dj + \frac{1}{N}\int_{\frac{r}{1-\alpha}}^N 0\times dj \\ &= \frac{1}{N}\left(\frac{r}{\alpha N}\frac{r}{1-\alpha} - \frac{1-\alpha}{\alpha N}\frac{1}{2}\left(\frac{r}{1-\alpha}\right)^2\right) = \frac{1}{2}\left(\frac{r}{N}\right)^2\frac{1}{\alpha\,(1-\alpha)},\end{aligned}$$

where limits on the definite integrals come from evaluting the *min* and *max* operators above inside the region $0 \leq r < \alpha N$.

Next, we can calculate the CDF in the second region – where $\alpha N \leq r < (1-\alpha) N$ – as:

$$\begin{aligned} R(r) &= \frac{1}{N}\int_0^{\frac{r-\alpha N}{1-\alpha}} 1 \times dj + \frac{1}{N}\int_{\frac{r-\alpha N}{1-\alpha}}^{\frac{r}{1-\alpha}} \frac{r-(1-\alpha)j}{\alpha N} dj + \frac{1}{N}\int_{\frac{r}{1-\alpha}}^{N} 0 \times dj \\ &= \frac{1}{N}\frac{r-\alpha N}{1-\alpha} + \frac{1}{N}\left(\frac{r}{\alpha N}\frac{r}{1-\alpha} - \frac{(1-\alpha)}{\alpha N}\frac{1}{2}\left(\frac{r}{1-\alpha}\right)^2\right) - \frac{1}{N}\left(\frac{r}{\alpha N}\frac{r-\alpha N}{1-\alpha} - \frac{(1-\alpha)}{\alpha N}\frac{1}{2}\left(\frac{r-\alpha N}{1-\alpha}\right)^2\right) \\ &= \frac{1}{N}\left(r^2 2\alpha N(1-\alpha)\right) + \frac{r}{(1-\alpha)N} - \frac{\alpha}{1-\alpha} - \frac{1}{2N}\left(\frac{r^2}{\alpha N(1-\alpha)} - \frac{\alpha N}{1-\alpha}\right) \\ &= \frac{r}{N}\frac{1}{1-\alpha} - \frac{1}{2}\frac{\alpha}{1-\alpha}. \end{aligned}$$

Finally, we can calculate the CDF in the third region – where $(1-\alpha) N \leq r \leq N$ – as:

$$\begin{aligned} R(r) &= \frac{1}{N}\int_0^{\frac{r-\alpha N}{1-\alpha}} 1 \times dj + \frac{1}{N}\int_{\frac{r-\alpha N}{1-\alpha}}^{N} \frac{r-(1-\alpha)j}{\alpha N} dj + \frac{1}{N}\int_{N}^{N} 0 \times dj \\ &= \frac{1}{N}\frac{r-\alpha N}{1-\alpha} + \frac{1}{N}\left(\frac{r}{\alpha} - \frac{N}{2}\frac{1-\alpha}{\alpha} - \frac{1}{2}\frac{r^2-\alpha^2N^2}{\alpha(1-\alpha)N}\right) \\ &= \frac{1}{2}\left(\frac{r}{N}\right)^2\frac{1}{\alpha(1-\alpha)} + \frac{r}{N}\frac{1}{\alpha(1-\alpha)} - \frac{1}{2}\left(\frac{1-\alpha}{\alpha} + \frac{\alpha}{1-\alpha}\right). \end{aligned}$$

Collecting these results yields the CDF for r in Equations (21).

B.1.4 Deriving the Aggregate Market Share s_j in Equation (26)

To derive the aggregate market share for a product j, we need to integrate spending shares across all households that buy j, where their spending shares are heterogeneous due to their receipt of an idiosyncratic taste shock $x_{i,j}$. Toward that end, we will use four new expressions:

1. Noting that $\tilde{\gamma}_{i,j} = G^{-1}\left(1 - R(r_{i,j})\right) = b\left(R(r_{i,j})\right)^{-\frac{1}{\theta}}$, we can substitute in to write the household's spending share on good j as:

$$s_{i,j} = P_i^{\sigma-1}\tilde{\gamma}_{i,j}^{\sigma-1} = (P_i b)^{\sigma-1}\left(R\left(r_{i,j}\right)\right)^{-\frac{\sigma-1}{\theta}},$$

   if $R(r_{i,j}) \leq |\Omega|/N$, and zero otherwise.

2. We need to solve for the cutoff rank value r^* so we can determine the worst idiosyncratic draw x_j^* for product j that still yields positive spending on that variety. *Focusing only on the first of the three regions* from the CDFs above, we have that the cutoff r^* satisfies $R(r^*) = |\Omega|/N$ or:

$$\frac{1}{2}\left(\frac{r^*}{N}\right)^2\frac{1}{\alpha(1-\alpha)} = \frac{|\Omega|}{N}.$$

   Solving the resulting quadratic equation for a positive root in $(0, N]$ leaves: $r^* = \left(2(1-\alpha)\alpha|\Omega|N\right)^{\frac{1}{2}}$.

3. We can then solve for the highest j good that experiences any positive consumption in the economy, j^*, as its rank value will equal r^* even when the idiosyncratic draw is the best possible case of $x = 0$. It will satisfy $r^* = (1-\alpha) j^*$, or $j^* = r^* / (1-\alpha) = \left(\frac{2\alpha|\Omega|N}{1-\alpha}\right)^{\frac{1}{2}}$.

4. Note that for a given good j, the worst possible idiosyncratic taste draw, x_j^* that yields positive consumption of j satisfies: $(1-\alpha) j + \alpha x_j^* = r^* = (1-\alpha) j^*$, or $x_j^* = \frac{1-\alpha}{\alpha}(j^* - j)$.

We can then solve for the aggregate expenditure share of good j as:

$$\begin{aligned}
s_j &= \frac{1}{\int_i E di} \int_i E s_{i,j} di = \frac{\eta+1}{2} N^{\frac{\eta-1}{2}} |\Omega|^{-\frac{\eta+1}{2}} \int_0^{x_j^*} \left(R\left((1-\alpha) j + \alpha x\right)\right)^{\frac{\eta-1}{2}} \frac{dx}{N} \\
&= \frac{\eta+1}{2} N^{\frac{\eta-1}{2}} |\Omega|^{-\frac{\eta+1}{2}} \int_0^{\frac{1-\alpha}{\alpha}(j^*-j)} \left(R\left((1-\alpha) j + \alpha x\right)\right)^{\frac{\eta-1}{2}} \frac{dx}{N} \\
&= \frac{\eta+1}{2} N^{\frac{\eta-1}{2}} |\Omega|^{-\frac{\eta+1}{2}} \int_0^{\frac{1-\alpha}{\alpha}(j^*-j)} \left(\frac{1}{2}\left(\frac{(1-\alpha) j + \alpha x}{N}\right)^2 \frac{1}{\alpha(1-\alpha)}\right)^{\frac{\eta-1}{2}} \frac{dx}{N} \\
&= \frac{\eta+1}{2} N^{\frac{\eta-1}{2}} |\Omega|^{-\frac{\eta+1}{2}} \int_0^{\frac{1-\alpha}{\alpha}(j^*-j)} 2^{\frac{1-\eta}{2}} N^{1-\eta} \left(\alpha(1-\alpha)\right)^{\frac{1-\eta}{2}} \left((1-\alpha) j + \alpha x\right)^{\eta-1} \frac{dx}{N} \\
&= (\eta+1) \left(2N|\Omega|\right)^{-\frac{\eta+1}{2}} \left(\alpha(1-\alpha)\right)^{\frac{1-\eta}{2}} \int_0^{\frac{1-\alpha}{\alpha}(j^*-j)} \left((1-\alpha) j + \alpha x\right)^{\eta-1} dx \\
&= (\eta+1) \left(2N|\Omega|\right)^{-\frac{\eta+1}{2}} \left(\alpha(1-\alpha)\right)^{\frac{1-\eta}{2}} \frac{1}{\alpha\eta} \left((1-\alpha) j + \alpha x\right)^{\eta} \Big|_0^{\frac{1-\alpha}{\alpha}(j^*-j)} \\
&= \frac{\eta+1}{\eta} \left(\left(\frac{2\alpha N|\Omega|}{1-\alpha}\right)^{\frac{1}{2}}\right)^{-\frac{\eta+1}{2}} \left((j^*)^{\eta} - j^e t a\right) \\
&= \frac{\eta+1}{\eta j^*} \left(1 - \left(\frac{j}{j^*}\right)^{\eta}\right).
\end{aligned}$$

B.1.5 Deriving the Aggregate Herfindahl $\mathcal{H}^{\text{Agg}}$ in Equation (27)

The Aggregate Herfindahl is calculated as:

$$\begin{aligned}
\mathcal{H}^{\text{Agg}} &= \int_0^{j^*} s_j^2 dj = \left(\frac{\eta+1}{\eta j^*}\right)^2 \int_0^{j^*} \left(1 - \left(\frac{j}{j^*}\right)^{\eta}\right)^2 dj \\
&= \left(\frac{\eta+1}{\eta j^*}\right)^2 \int_0^{j^*} \left(1 - 2\left(\frac{j}{j^*}\right)^{\eta} + \left(\frac{j}{j^*}\right)^{2\eta}\right) dj \\
&= \left(\frac{\eta+1}{\eta j^*}\right)^2 \left[j - 2\left(\frac{1}{j^*}\right)^{\eta} \frac{j^{\eta+1}}{\eta+1} + \left(\frac{1}{j^*}\right)^{2\eta} \frac{j^{2\eta+1}}{2\eta+1}\right]_0^{j^*} \\
&= \left(\frac{\eta+1}{\eta j^*}\right)^2 \left[j^* - 2\frac{j^*}{\eta+1} + \frac{j^*}{2\eta+1}\right] \\
&= \frac{2(\eta+1)}{2\eta+1} \frac{1}{j^*} \\
&= \frac{2(\eta+1)}{2\eta+1} \left(\frac{1}{2|\Omega|\tilde{N}}\right)^{\frac{1}{2}},
\end{aligned}$$

where we define $\tilde{N} = \alpha N / (1-\alpha)$.

B.1.6 Deriving Elasticities ε_j and Markups μ_j and μ^{Agg} in Equations (36), (37), and (38)

To solve for the price elasticity of aggregate demand for product j, we start by expressing its total sales as the integral of each household's spending on j, taken over all households:

$$s_j = \frac{1}{N}\int_0^{x_j^*} s_{i_x,j} dx, \tag{A2}$$

where we use the notation $s_{i_x,j}$ to denote the spending share of a household with taste draw on product j equal to x. We take the partial derivative of s_j in equation (A2) with respect to p_j to get:

$$\frac{\partial s_j}{\partial p_j} = \frac{1}{N}\left(\int_0^{x_j^*} \frac{\partial s_{i_x,j}}{\partial p_j} dx + s_{i_{x_j^*},j}\frac{\partial x_j^*}{\partial p_j}\right), \tag{A3}$$

where the right hand side of equation (A3) follows from Leibniz's rule. The first term can be solved using equation (22) as:

$$\frac{\partial s_{i_x,j}}{\partial p_j} = \frac{\partial P^{\sigma-1} p_j^{1-\sigma} \gamma_{i,j}^{\sigma-1}}{\partial p_j} = (1-\sigma)\frac{s_{i_x,j}}{p_j}, \tag{A4}$$

where we take the aggregate price index P as fixed. Moving on to the second term, we can evaluate equation (22) at the marginal household with taste x_j^* to get:

$$s_{i_{x_j^*},j}\frac{\partial x_j^*}{\partial p_j} = \frac{\eta+1}{2} N^{\frac{\eta-1}{2}} |\Omega|^{-\frac{\eta+1}{2}} \left(R\left(r^*\right)\right)^{\frac{\eta-1}{2}} \frac{\partial x_j^*}{\partial p_j}. \tag{A5}$$

Substituting equations (A4) and (A5) back into equation (A3), we get:

$$\begin{aligned}
\frac{\partial s_j}{\partial p_j} &= (1-\sigma)\frac{1}{p_j}\frac{1}{N}\int_0^{x_j^*} s_{i_x,j} dx + \frac{1}{N} s_{i_{x_j^*},j}\frac{\partial x_j^*}{\partial p_j} \\
&= (1-\sigma)\frac{s_j}{p_j} + \frac{1}{N}\frac{\eta+1}{2}\left(|\Omega|\right)^{-\frac{1+\eta}{2}} N^{\frac{\eta-1}{2}} \left(R\left(r^*\right)\right)^{\frac{\eta-1}{2}} \frac{\partial x_j^*}{\partial p_j} \\
&= (1-\sigma)\frac{s_j}{p_j} + \frac{1}{N}\frac{\eta+1}{2}\left(|\Omega|\right)^{-\frac{1+\eta}{2}} N^{\frac{\eta-1}{2}} \left(\frac{1-\alpha}{2\alpha N^2}\left(j^*\right)\right)^{\frac{\eta-1}{2}} \frac{\partial x_j^*}{\partial p_j} \\
&= (1-\sigma)\frac{s_j}{p_j} + \frac{\eta+1}{2N|\Omega|}\frac{\partial x_j^*}{\partial p_j}.
\end{aligned} \tag{A6}$$

To approximate $\partial x_j^* / \partial p_j$, we start with the relationship:

$$\frac{1}{2}\left(\frac{r_{i,j}}{N}\right)^2 \frac{1}{\alpha(1-\alpha)} = R\left((1-\alpha)j + \alpha x_{i,j}\right) = 1 - G\left(\frac{\gamma_{i,j}}{p_j}\right) = b^\theta \gamma_{i,j}^{-\theta} p_j^\theta, \tag{A7}$$

and differentiate to yield:

$$\frac{r_{i,j}}{1-\alpha}\frac{1}{N^2}\frac{\partial x_{i,j}}{\partial p_j} = \theta b^\theta \gamma_{i,j}^{-\theta} p_j^{\theta-1}, \tag{A8}$$

where we've substituted $\partial r_{i,j}/\partial x_{i,j} = \alpha$. We then evaluate equation (A8) at $r_{i,j} = r^*$ and $\tilde{\gamma}_{i,j} = \tilde{\gamma}^*$ using equation (23) and add a minus sign to reflect the fact that increase in the price of good j should reduce the set of households purchasing that good, to get:

$$\begin{aligned}
\frac{\partial x_j^*}{\partial p_j} &= -\theta b^\theta \gamma_{i,j}^{-\theta} p_j^{\theta-1} N^2 \frac{1-\alpha}{r^*} \\
&= -\theta b^\theta \left(\tilde{\gamma}^*\right)^{-\theta} p_j^{-1} N^2 \frac{1}{j^*} \\
&= -\theta \frac{|\Omega|}{N} p_j^{-1} N^2 \frac{1}{j^*} \\
&= -\theta |\Omega| N p_j^{-1} \frac{1}{j^*} \\
&= -\frac{\theta}{j^*} |\Omega| N \frac{1}{p_j}.
\end{aligned}$$

Inserting this into equation (A6), we have:

$$\begin{aligned}
\frac{\partial s_j}{\partial p_j} &= (1-\sigma)\frac{s_j}{p_j} + \frac{\eta+1}{2N|\Omega|}\frac{\partial x_j^*}{\partial p_j} \\
&= (1-\sigma)\frac{s_j}{p_j} - \frac{\eta+1}{2N|\Omega|}\frac{\theta}{j^*}|\Omega|N\frac{1}{p_j} \\
&= (1-\sigma)\frac{s_j}{p_j} - \frac{\eta+1}{j^*}\frac{\theta}{2}\frac{1}{p_j} \\
&= (1-\sigma)\frac{s_j}{p_j} - \frac{\eta\theta}{2\left(1-\left(\frac{j}{j^*}\right)^\eta\right)}\frac{s_j}{p_j},
\end{aligned} \tag{A9}$$

where the last substitution uses the definition of product share in equation (26). Equation (35) implies that product j's price elasticity of demand ε_j can be written as:

$$\varepsilon_j = 1 - \frac{\partial s_j}{\partial p_j}\frac{p_j}{s_j} = \underbrace{\sigma}_{\text{Intensive Margin}} + \underbrace{\left(1-\left(\frac{j}{j^*}\right)^\eta\right)^{-1}\left[\theta/2-(\sigma-1)\right]}_{\text{Extensive Margin}} > \sigma. \tag{A10}$$

The markup μ_j then be written (in gross terms) as:

$$\mu_j = \frac{\varepsilon_j}{\varepsilon_j - 1} = \frac{\sigma + \frac{\theta(\eta+1)}{2j^* s_j}}{\sigma + \frac{\theta(\eta+1)}{2j^* s_j} - 1}. \tag{A11}$$

The aggregate markup μ^{Agg} is equal to the ratio of aggregate sales to aggregate costs. Using

equations (26) and (A10), it can be written as:

$$\mu^{\text{Agg}} = \frac{\int_0^{j^*} s_j dj}{\int_0^{j^*} s_j \frac{\varepsilon_j - 1}{\varepsilon_j} dj} = \left[\frac{\theta + (\sigma - 1)^2}{\sigma^2} - \frac{1}{2} \frac{\eta \theta^2}{\sigma^2} \left(\frac{\eta + 1}{2 + \theta} \right) \times {}_2F_1 \left(1, \frac{1}{\eta}; 1 + \frac{1}{\eta}; \frac{2\sigma}{2 + \theta} \right) \right]^{-1}, \tag{A12}$$

where we've used Mathematica to evaluate and simplify this final expression.

B.1.7 Testable Predictions at Household Level

We now use the model to derive a relationship between a product's share of a household's total spending and its rank in that household's bundle, given the total number of goods in the bundle. In particular, we show how to convert our model which assumes a continuum of products are purchased into a discrete interpretation which can be tested directly in the data.

We start with expression (17) from the paper, which states that for household i that ranks product $k \leq |\Omega|$, k's spending share will equal:

$$\begin{aligned} s_{i,k} &= (P \tilde{\gamma}_{i,k})^{\sigma - 1} && \text{(A13)} \\ &= \left(\left(1 + \frac{1 - \sigma}{\theta} \right)^{\frac{1}{\sigma - 1}} b^{-1} \left(|\Omega_i| \right)^{\frac{1}{1 - \sigma}} \left(\frac{|\Omega_i|}{N} \right)^{\frac{1}{\theta}} \tilde{\gamma}_{i,k} \right)^{\sigma - 1} && \text{(A14)} \\ &= \left(1 + \frac{1 - \sigma}{\theta} \right) \left(|\Omega_i| \right)^{-1} b^{1 - \sigma} \left(\frac{|\Omega_i|}{N} \right)^{\frac{\sigma - 1}{\theta}} \left(\tilde{\gamma}_{i,k} \right)^{\sigma - 1}, && \text{(A15)} \end{aligned}$$

where the second line substitutes in the definition of P from equation (14) in the paper.

Define the percentile $p_{i,k}$ of good k in i's bundle as:

$$p_{i,k} = \frac{1 - G\left(\tilde{\gamma} \right)}{1 - G\left(\tilde{\gamma}^* \right)} = \frac{N}{|\Omega_i|} \left(\frac{\tilde{\gamma}_{i,k}}{b} \right)^{-\theta}, \tag{A16}$$

where a value of $p_{i,k} = 0.10$, for example, means variety k is at the 10th percentile out of all those that are consumed and is thus preferred to 90 percent of the goods consumed by that household. Substituting, this then allows us to write the expenditure share of a good as a decreasing function of its percentile (ranging from 0 to 1) in terms of how preferred it is among the goods that are consumed:

$$s_{i,k} = \frac{\eta + 1}{2 |\Omega_i|} p_{i,k}^{\frac{\eta - 1}{2}}. \tag{A17}$$

Note that integrating over all consumed goods k, the shares sum to one:

$$\int_0^1 \frac{\eta + 1}{2 |\Omega_i|} p_{i,k}^{\frac{\eta - 1}{2}} |\Omega_i| dp_{i,k} = 1. \tag{A18}$$

However, our data is discrete. One might be tempted to simply substitute $p_{i,k} = 1/|\Omega_i|$ for the highest

share good in our data, and $p_{i,k} = 2/|\Omega_i|$ for the second, and so-on, and then estimate the equation:

$$\ln(s_{i,k}) = \ln\left(\frac{\eta+1}{2|\Omega_i|}\right) + \frac{\eta-1}{2}\ln(p_{i,k}) + \epsilon_{i,k}, \tag{A19}$$

constraining the slope to be $|\Omega_i|$ times the intercept and then comparing data relative to model predictions for various values of Ω_i. The problem with this is that since our data is discrete, the equation does not exactly hold. For example, if there are 8 goods, using the formula with $p = \{1/8, 2/8, ..., 1\}$, does not imply shares that sum to one.

We therefore derive an analog of equation (A17) for a world with discrete goods. In particular, define $s_{i,K}^{|\Omega_i|}$ for $K \in \{1, 2, ..., |\Omega_i|\}$ as the combined spending share of all varieties in the span of $[(K-1)/|\Omega_i|, K/|\Omega_i|]$. We then associate $s_{i,K}^{|\Omega_i|}$ with the spending share of good with rank K in the data. We calculate:

$$s_{i,K}^{|\Omega_i|} = \int_{(K-1)/|\Omega_i|}^{K/|\Omega_i|} \frac{\eta+1}{2|\Omega_i|} p_{i,k}^{\frac{\eta-1}{2}} |\Omega_i| dp_{i,k} \tag{A20}$$

$$= (|\Omega_i|)^{-\frac{\eta+1}{2}} \left[K^{\frac{\eta+1}{2}} - (K-1)^{\frac{\eta+1}{2}}\right]. \tag{A21}$$

These shares then are functions of both $|\Omega_i|$ and η. For example, the simplest case is the share spent on the most preferred good (i.e. $K = 1$) which can be written as:

$$s_{i,1}^{|\Omega_i|} = (|\Omega_i|)^{-\frac{\eta+1}{2}}. \tag{A22}$$

The other cases are similar but the bracketed term in equation (A21) does not simplify. Figure 7 in the paper then demonstrates that this theoretical relationship (A21) aligns closely with the spending patterns in our data.

B.1.8 Relationship to Feenstra (1994)

In this subsection, we connect the standard Feenstra (1994) correction with the price index in our environment. Let F_t denote the Feenstra correction at t, defined as:

$$F_t = \frac{P_t / P_{t-1}}{P_t^{\text{Conventional}} / P_{t-1}^{\text{Conventional}}}, \tag{A23}$$

where P_t is the exact price index at t corresponding to demand over varieties j with a constant elasticity of substitution (CES) and where $P_t^{\text{Conventional}}$ is a "conventional" price index that is measured entirely using varieties that are consumed in two consecutive periods:

$$\frac{P_t^{\text{Conventional}}}{P_{t-1}^{\text{Conventional}}} = \prod_j \left(p_{j,t}/p_{j,t-1}\right)^{\omega_{j,t}}, \tag{A24}$$

where $p_{j,t}$ is the price of good j at t and the weights $\omega_{j,t}$ are Sato-Vartia weights and are functions of the expenditure shares of good j in t and $t-1$.

There are four key quantities and one key parameter in the Feenstra correction. It equals:

$$F_t = \left(\frac{F_{1,t}/F_{2,t}}{F_{3,t}/F_{4,t}} \right)^{-\frac{1}{\sigma-1}}. \tag{A25}$$

$F_{1,t}$ is the spending at t, $F_{2,t}$ is the spending at t on goods that were also purchased at $t-1$, $F_{3,t}$ is the spending at $t-1$, and $F_{4,t}$ is the spending at $t-1$ on goods that will eventually be purchased at t.

To express the Feenstra correction at the household level in our environment, consider an increase in N from N_{t-1} to N_t that increases Ω_{t-1} to Ω_t. Note that $N_t > N_{t-1}$ causes $|\Omega_t| > |\Omega_{t-1}|$ and $\tilde{\gamma}_t^* > \tilde{\gamma}_{t-1}^*$, because the increase in N exceeds the increase in $|\Omega|$ (i.e. $|\Omega|/N$ goes down). So in period t, all products with values of $\tilde{\gamma} \in \left(\tilde{\gamma}_{t-1}^*, \tilde{\gamma}_t^*\right)$ that were consumed in $t-1$ are dropped, and every value of $\tilde{\gamma}$ that was purchased in t was also purchased in $t-1$. However, since the density of varieties with any given $\tilde{\gamma}$ value is N_t/N_{t-1} higher, we have:

$$F_{2,t} = E\frac{N_{t-1}}{N_t}. \tag{A26}$$

For $F_{4.t}$, we simply want to calculate spending during $t-1$ on all $\tilde{\gamma}$ values greater than $\tilde{\gamma}_t^*$, since those varieties would also have been consumed at t:

$$\begin{aligned}
F_{4,t} &= EN_{t-1}\int_{\tilde{\gamma}_t^*}^{\infty} (P_{t-1})^{\sigma-1} (y)^{\sigma-1} dG(y) && \text{(A27)}\\
&= EN_{t-1} (P_{t-1})^{\sigma-1} \int_{\tilde{\gamma}_t^*}^{\infty} (y)^{\sigma-1} dG(y) && \text{(A28)}\\
&= EN_{t-1} (P_{t-1})^{\sigma-1} \left(\frac{\theta b^\theta}{\theta-\sigma+1} \right) (\tilde{\gamma}_t^*)^{\sigma-1-\theta} && \text{(A29)}\\
&= Eb^{\theta+1-\sigma} \left(\frac{|\Omega_{t-1}|}{N_{t-1}} \right)^{\frac{\sigma-1-\theta}{\theta}} (\tilde{\gamma}_t^*)^{\sigma-1-\theta} && \text{(A30)}\\
&= E \left(\frac{|\Omega_t|/|\Omega_{t-1}|}{N_t/N_{t-1}} \right)^{\frac{\theta+1-\sigma}{\theta}} && \text{(A31)}
\end{aligned}$$

Remember that the terms F_1 and F_3 just capture spending in t and $t-1$, so they are equal: $F_{1,t} = F_{3,t} = E$.

Putting this all together into equation (A25), we have

$$F_t = \left(\frac{F_{4,t}}{F_{2,t}} \right)^{-\frac{1}{\sigma-1}} = \left(\frac{\left(\frac{|\Omega_t|/|\Omega_{t-1}|}{N_t/N_{t-1}} \right)^{\frac{\theta+1-\sigma}{\theta}}}{\frac{N_{t-1}}{N_t}} \right)^{-\frac{1}{\sigma-1}}$$

$$
\begin{aligned}
&= \left(\frac{|\Omega_t|}{|\Omega_{t-1}|}\left(\frac{|\Omega_t|/|\Omega_{t-1}|}{N_t/N_{t-1}}\right)^{\frac{1-\sigma}{\theta}}\right)^{-\frac{1}{\sigma-1}} \\
&= \underbrace{\left(\frac{|\Omega_t|}{|\Omega_{t-1}|}\right)^{\frac{1}{1-\sigma}}}_{\text{Change in Variety Effects}} \times \underbrace{\left(\frac{|\Omega_t|/N_t}{|\Omega_{t-1}|/N_{t-1}}\right)^{\frac{1}{\theta}}}_{\text{Change in Selection Effects}}, \qquad \text{(A32)}
\end{aligned}
$$

where we can see that the first terms equal the growth from period $t-1$ to period t of the "Variety Effects" and "Selection Effects" terms derived in the ideal price index in the main paper. The "Average Price" term does not appear as the experiment of increasing N should leave unchanged the full price distribution and therefore there'd be no growth in $P^{\text{Conventional}}$.

The analysis is largely similar if we consider what one would obtain were they to mistakenly assume behavior in our environment was generated from a representative household and calculated a Feenstra correction using aggregate spending. As before, we'd have that $F_{1,t} = F_{3,t} = E$, as well as that $F_{2,t} = E \times N_{t-1}/N_t$. But now, we solve for $F_{4,t}$ by integrating the formula for the aggregate expenditure share – equation (26) – (times total expenditures) in period $t-1$ from the top good in terms of aggregate spending ($j = 0$) to the cutoff value of $\tilde{\gamma}_t^*$ which obtains at period t, which equals $j_t^* \times N_{t-1}/N_t$:

$$
\begin{aligned}
F_{4,t} &= E\int_{j=0}^{j_t^*\frac{N_{t-1}}{N_t}} s_j dj = E\int_{j=0}^{j_t^*\frac{N_{t-1}}{N_t}} \frac{\eta+1}{\eta j_{t-1}^*}\left(1-\left(\frac{j}{j_{t-1}^*}\right)^{\eta}\right)dj \\
&= E\frac{\eta+1}{\eta j_{t-1}^*}\left[j - \frac{1}{\eta+1}j^{\eta+1}\left(j_{t-1}^*\right)^{-\eta}\right]_0^{j_t^*\frac{N_{t-1}}{N_t}} \\
&= E\frac{\eta+1}{\eta}\left(\frac{j_t^*/N_t}{j_{t-1}^*/N_{t-1}} - \frac{1}{\eta+1}\left(\frac{j_t^*/N_t}{j_{t-1}^*/N_{t-1}}\right)^{\eta+1}\right) \\
&= E\frac{\eta+1}{\eta}\left(\left(\frac{|\Omega_t|/N_t}{|\Omega_{t-1}|/N_{t-1}}\right)^{\frac{1}{2}} - \frac{1}{\eta+1}\left(\frac{|\Omega_t|/N_t}{|\Omega_{t-1}|/N_{t-1}}\right)^{\frac{\eta+1}{2}}\right), \qquad \text{(A33)}
\end{aligned}
$$

where the last line substitutes in the relationship in equation (24). This then leads to the Feenstra correction one would make if they incorrectly assumed the data were generated by a single household:

$$
F_t = \frac{\eta+1}{\eta}\frac{N_t}{N_{t-1}}\left(\left(\frac{|\Omega_t|/N_t}{|\Omega_{t-1}|/N_{t-1}}\right)^{\frac{1}{2}} - \frac{1}{\eta+1}\left(\frac{|\Omega_t|/N_t}{|\Omega_{t-1}|/N_{t-1}}\right)^{\frac{\eta+1}{2}}\right)^{-\frac{1}{\sigma-1}}. \qquad \text{(A34)}
$$

It is clear that, plugging the values for $|\Omega|$, N, and the other structural parameters into equations (A32) and (A34) will yield different answers. As discussed in the main text, equation (A32) is relevant for the household's welfare in our model, but equation (A34) is not since our model does not admit a representative agent with CES preferences.

B.2 Model Simulation Results

In this section, we explore numerical simulations of our model to test the validity of our elasticity approximation as well as to explore how restrictive the assumption of a stable distribution of Pareto taste-adjusted prices is for our conclusion.

We simulate a discrete approximation to the main model in the paper by drawing a large random vector $\tilde{\gamma}^{rand}$ of price-adjusted tastes from a Pareto distribution for a large sample of households, using the same parameters as our baseline model.[12] Our random sample uses 2.25 million draws for each of 20,000 households since we are trying to approximate a continuum of products from [0,N]. However, rather than using analytical formulas to calculate household market shares, for each household we instead keep the random set $\Omega = \tilde{\gamma}^{rand} > \tilde{\gamma}^{*}$ and then calculate a numerical price index directly from $P = \left(\int_{k\in\Omega_i} (\tilde{\gamma}_{i,k})^{\sigma-1}\, dk\right)^{\frac{1}{1-\sigma}}$ and then compute market shares from equation 17. These formulas hold for arbitrary distributions of taste, so even though we still simulate the taste draws from a Pareto distribution, in this simulation we are using no analytical results that rely on this assumption, which also means that we can also perform a similar procedure even if tastes do not follow a Pareto distribution.

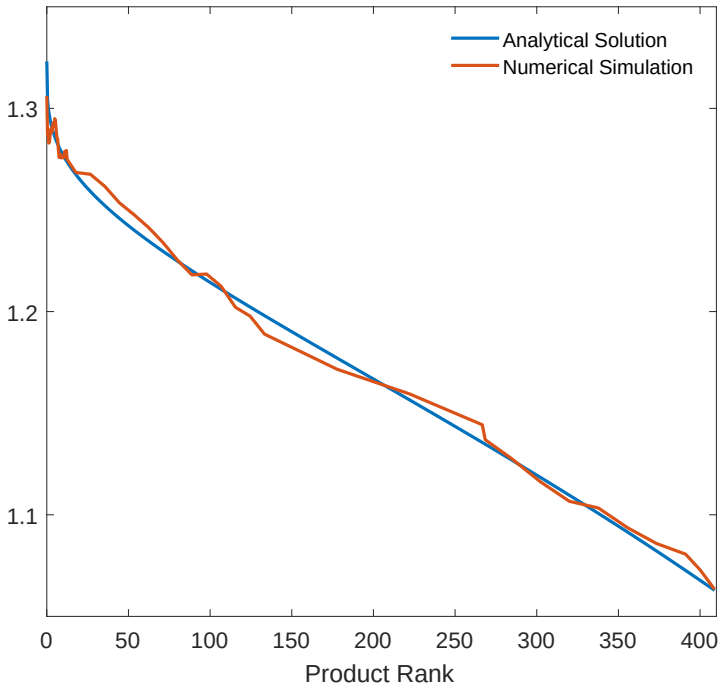
In order to get aggregate market shares, we must identify the particular products that each household consumes. In order to do so, we use our rank function equation 20 with a random uniform draw to compute for each household, the aggregate ranking of each of the 2.25 million possible products in [0,N] and then compute household $i's$ particular idiosyncratic rank for each of the 2.25 million j products. We then sort $\tilde{\gamma}^{rand}$ and map the highest value to $r_{i,j} = 0$, the second highest value to $r_{i,j} = 1$ and so on to $r_{i,j} = 2.25$ million. Finally, since for each $r_{i,j}$, we know the value of j, this means that we then know household i's taste draw and resulting individual spending for each aggregate product j. For example, the households highest $\tilde{\gamma}^{rand}$ draw will always map to their $r_{i,j} = 0$, but the corresponding aggregate j which household 1 ranks highest might be $j = 0$, the j which household 2 ranks highest might be $j = 2043$, and the j which household 3 ranks highest might be $j = 17$. Once we have these household specific spending shares for each product j, we can then numerically add up total spending on each product j to calculate aggregate market shares.

Since these are computed entirely numerically, they do not rely on any of our closed form solutions for aggregate market shares and are thus again valid even under departures from the Pareto distribution. As we note in 4.7, our analytical market shares are only valid under the Pareto distribution so we must approximate the elasticity of demand by modeling a price change as a switch with another product in the aggregate ranking. Since these numerical results do not rely on the Pareto distribution, we can use this numerical model to simulate the aggregate elasticity of demand and resulting markup for a product j by just raising all households' random taste draw for that product by a small amount.

[12] Note that in this numerical exercises we are not restricted to this distribution. We could instead assume a Pareto distribution on tastes given arbitrary prices or remove the Pareto assumption entirely. However, only the case where taste-adjusted prices are initially Pareto can be used to compare to our analytical results, so we focus on this numerical case.

Note that calculating elasticities for each j requires re-simulating a new set of aggregate market shares. For these sample sizes, computing an elasticity for a single j requires roughly 2 hours of computational time, so it is infeasible to simulate the elasticity of demand for all 2.5 million products. Instead, we compute the elasticity of demand and implied markups for 50 different values of j distributed throughout the product space. Figure A20 compares this simulated markup to our analytical approximation and shows that the analytical approach produces essentially identical results (noting that there is still obvious numerical simulation error even with these large sample sizes).

Figure A20: Simulated vs. Analytical Approximation for Markup



As stressed throughout the paper, our analytical derivations and implications of changes in N are only valid under the assumption that the distribution of price-adjusted tastes continues to follow a Pareto distribution as we vary N. If markups were fixed for all products, then assuming that the distribution of price-adjusted tastes is held fixed as N varies would be a natural benchmark. However, our model instead implies that optimal markups do vary across products, and that the markups for individual products change as we vary N. This implies that if household tastes for products and their marginal costs held fixed, but we allow prices to change along with optimal markups when N changes, then there will necessarily be a violation of the assumed Pareto distribution. Since all of our analytical results assume the Pareto distribution of price-adjusted tastes, this means that our analytical comparative statics to changes in N and F which induce changes in product markups are technically comparative statics in response to these parameter changes plus whatever implicit changes in household tastes (or marginal costs) are necessary to preserve a Pareto distribution of price-adjusted tastes

after markups adjust. In practice, high turnover means that the set of products purchased in 2004 and in 2016 is mostly disjoint, so one can primarily interpret these as taste shifts for new products rather than taste changes for existing products.[13] However, if the required taste shifts necessary to maintain the Pareto distribution under our counterfactuals were substantial, then this would potentially substantively change the interpretation of the welfare effects of changes in N.

However, we now use our numerical model to show that even though there are indeed implicit taste changes necessary to maintain the Pareto distribution as N changes, in practice these required taste changes are quantitatively small and actually work against our conclusion that N is welfare improving. We thus conclude that even though this is a large potential issue for the interpretation of our comparative statics, it is of little quantitative importance in practice. Specifically, we perform the following exercise: For the initial value of N in 2004, we simulate our numerical model exactly as described above. Given household i's resulting distribution of tastes for all j products $\tilde{\gamma}_{i,j}^{rand}$, we can then compute a households actual (non-price adjusted) taste for product j $\gamma_{i,j}^{rand} = \tilde{\gamma}_{i,j}^{rand}\mu_j$ using the analytical formula for μ_j from Section 4.7.[14] Note that as we explore above, even though our numerical model does not otherwise rely on analytical results, this analytical formula for the markup is valid since we are drawing the numerical distribution of price-adjusted tastes in the model from a Pareto distribution.

We then increase N in the model but hold the particular random realizations of $\tilde{\gamma}_{i,j}^{rand}$ exactly fixed in the new simulation. Thus, by assumption, the values of *price-adjusted* tastes will be identical in the two simulations. However, as N increases, the function μ_j and resulting prices will change. If price-adjusted tastes are fixed by assumption, but prices change then household tastes must change.[15].

How large are the required taste changes necessary to maintain an identical realization from a Pareto distribution of price-adjusted tastes as N increases? Figure A21 shows that these changes are small. The left panel plots the implied taste draws as a function of initial aggregate product rank j for a fixed household before and after a 70% increase in N.[16] Clearly the increase in N induces some implied changes in tastes in order to maintain the Pareto distribution for price-adjusted tastes, but it is also clear that the requisite taste changes are small. The right panel of the plot shows a scatter plot of the realizations of taste before and after the increase in N. Overall the R^2 is above 0.999, so there is an almost perfect correlation of tastes under the two scenarios. In order to maintain an identical distribution of price-adjusted tastes, there is a modest *decline* in the implied average taste when N increases, which lowers implied welfare by roughly 1.3%. This occurs because as N increases, markups for incumbent products decline, which makes price fall and thus taste/price rise. In order to

[13] Only 13.2% of UPCs purchased in 2004 are still purchased in 2016.

[14] For notational simplicity, we assume that marginal cost is 1 for all products. More generally this approach actually recovers the distribution of marginal cost adjusted tastes. As long as we assume marginal cost is constant as we vary N, one can interpret changes in taste and changes in marginal cost adjusted taste equivalently so these are equivalent exercises.

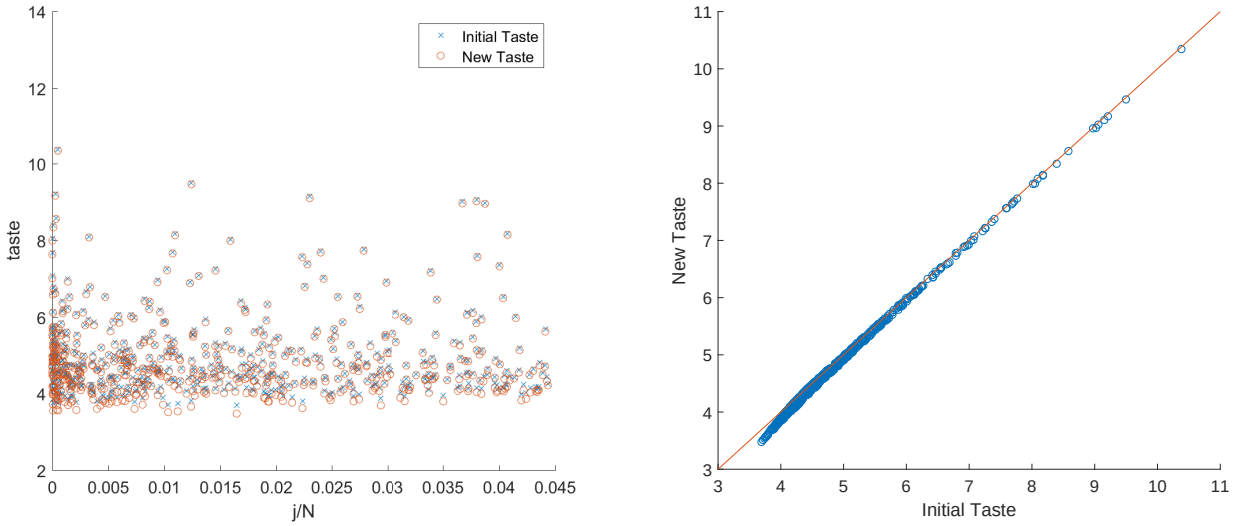
[15] Further, since markup changes are a monotonic function of j but individual rankings of the j products are non-monotonic when $\alpha > 1$, these price changes will be non-monotonic over individual households' consumption baskets.

[16] Here we focus on products which are consumed in both scenarios so that such taste comparisons are relevant.

maintain a constant taste/price for that product, this means taste for those products must decline.

Figure A21: Household Taste Changes Required to Maintain Pareto with N increase

(a): Initial and New Tastes by Aggregate Product Rank **(b):** Initial vs. New Tastes



However, the welfare conclusion in the body of the paper under the assumed constant Pareto distribution of price-adjusted tastes is that an increase in N of 70% *raises* welfare by roughly 9.5%. The numerical results above show that that those welfare results are only valid if there is also a simultaneous modest decline in non-price adjusted tastes when N rises, suggesting that if one instead held tastes fixed when increasing N and departed from Pareto, the welfare increase would be slightly stronger. While such an exercise could potentially be performed numerically, it would require solving for the entire equilibrium distribution of the elasticity of demand and resulting markups numerically. As discussed above, the numerical calculation of the elasticity of demand (even for a single product in partial equilibrium) is very computationally costly.

Finally, we use this simulated model to also explore the role of potential measurement error in driving concentration trends. Although A.2 shows that the Nielsen data tracks aggregate spending measures fairly closely, the declining within-household spending patterns suggest there may be some role for attrition related measurement error across time. Furthermore, even though households are supposed to report online purchases and that Figure A1 shows that online spending is relatively unimportant for these sectors, it is possible that under-reported online spending might also drive increasing measurement error across time.

While it is difficult to analytically characterize the role of various forms of measurement error for concentration trends, we follow the indirect inference approach in Berger and Vavra (2015) and Berger and Vavra (2019) and *simulate* various flexible forms of measurement error in the numerical version of our model under the assumption that all other model parameters are held fixed. Specifically, we

simulate the discrete version of our model and separately consider the effects of measurement error on household and aggregate concentration. We focus primarily on measurement error arising from failing to report transactions entirely rather than from misreporting the size of a transaction, since the former is much more likely given the structure of the Homescan data collection. We consider three types of potential under-reporting encompassing various different extremes: 1) households failing to report some randomly chosen purchases, 2) households failing to report their smallest purchases and 3) households failing to report their largest transactions. Overall, we find that while measurement error can change both household and aggregate concentration, it pushes both household and aggregate concentration in the same direction and so is unlikely to be an important explanation for the observed rise in niche consumption. Unsurprisingly, the first and second form of measurement error raise both household and aggregate concentration while the third form of measurement error instead lowers both concentration measures. Furthermore, the second form of measurement error seems most plausible given the nature of the Nielsen data, since a household might fail to report a small one-off purchase which is likely to be a small share of that household's annual spending but is unlikely to consistently fail to report large, regular purchases that are likely to be a large share of annual spending. Since the third form of measurement error is especially unlikely, this means that measurement error is then also quite unlikely to explain a decline in aggregate concentration. A decline in aggregate concentration with flat household concentration would generally be sufficient to infer an increase in N. Overall, these simulation results strongly suggest that measurement error does not drive the rise of niche consumption.

B.3 Specification with Linear Demand

In this subsection, we sketch a comparable setup to our baseline but using a linear demand system, following Melitz and Ottaviano (2008). We include key analytical results, such as derivations of Household and Aggregate Herfindhals, but additional details are available on request. We owe particular thanks to Levi Crews and Agustin Gutierrez for their outstanding research assistance in deriving the below expressions.

B.3.1 Setup

We assume household preferences are defined over a continuum of differentiated varieties indexed by $k \in (0, N]$. All consumers share the same utility function given by

$$U_i = \beta \int_{k\in(0,N]} \gamma_{i,k} C_{i,k} dk - \frac{1}{2}\sigma \int_{k\in(0,N]} (\gamma_{i,k} C_{i,k})^2 dk - \frac{1}{2}\eta \left(\int_{k\in(0,N]} \gamma_{i,k} C_{i,k} dk \right)^2,$$

where β, σ, η, and all preference shifters $\gamma_{i,k}$ are all non-negative. We impose an additional cost of consuming differentiated varieties, which we assume takes an exponential form in the measure of

consumed varieties. Specifically, household i pays a cost $F \times (|\Omega_i|)^{\epsilon}$ in units of the numeraire, where Ω_i is the set of differentiated varieties consumed in a positive amount by household i and F is non-negative. Household i therefore solves:

$$\max_{\{C_{i,k}\}} \; U_i \quad \text{s.t.} \quad \int_{k\in(0,N]} p_{i,k} C_{i,k} dk + F \times (|\Omega_i|)^{\epsilon} \leq E.$$

We continue assume that price-adjusted tastes for each $k \in (0, N]$ are distributed Pareto for each household i with shape θ and support $[b, \infty)$ with $b > 0$. Because there is a continuum of varieties, each household faces the same set of taste-adjusted prices, though as before, their ranking of varieties within that set may differ. It can be shown that there exists a unique optimal measure of consumed varieties if and only if $\frac{\beta}{b}\left(\frac{\theta}{\theta+2}\right) N > E - F \times N^{\epsilon}$. In the expressions that follow, assume this condition holds.

B.3.2 Key Expressions

Household expenditure shares in this environment can be written in closed-form as:

$$s_{i,k} = \frac{1}{|\Omega_i|}\left(\frac{(\theta+1)\,(\theta+2)}{\theta}\right)\left(1 - \frac{\tilde{\gamma}^*}{\tilde{\gamma}_{i,k}}\right)\frac{\tilde{\gamma}^*}{\tilde{\gamma}_{i,k}},$$

where $\tilde{\gamma}^* = |\Omega|^{-\frac{1}{\theta}} N^{\frac{1}{\theta}} b$. This then allows us to calculate and express the Household Herfindahl as:

$$\mathcal{H}^{HH} = \left[\frac{2(\theta+2)(\theta+1)^2}{\theta(\theta+3)(\theta+4)}\right]\frac{1}{|\Omega|}.$$

To move to the Aggregate Herfindahl, we impose the same form for the rank function as we used in the main analysis and obtain an expression for the aggregate share of product j:

$$s_j = \left[\frac{(\theta+1)(\theta+2)}{\theta}\right]\left[\frac{2(1-\alpha)}{\alpha N|\Omega|}\right]^{\frac{1}{2}}\left\{\frac{\theta}{\theta+2}\left[1 - \left(\frac{j}{j^*}\right)^{\frac{2}{\theta}+1}\right] - \frac{\theta}{\theta+4}\left[1 - \left(\frac{j}{j^*}\right)^{\frac{4}{\theta}+1}\right]\right\}.$$

This leads to the expression for the Aggregate Herfindahl:

$$\mathcal{H}^{\text{Agg}} = \frac{8}{3}\left[\frac{(\theta+1)(7\theta+12)}{(3\theta+4)(3\theta+8)}\right]\left[\frac{1}{2\tilde{N}|\Omega|}\right]^{\frac{1}{2}}.$$

B.3.3 Comparison to the CES Case

When we confront these expressions with the data to extract changes in θ and $\tilde{N}$, we find very similar results as what we found in the baseline CES case. In particular, the implied $\tilde{N}$ increases by roughly 70-80 percent, while the value for θ declines very slightly.

The implied value for θ, however, is much lower in the linear demand model than in the CES

model. Though there is not a clear empirical benchmark for the value of θ, the value strikes as as less plausible than that from the CES case. Further, the aggregate market share distribution for such low values of θ do not conform as well with the CDF of market shares in our Nielsen data, as plotted in Figure 8. As a result, we use the CES specification as our benchmark.