Retail Globalization, Households' Diets and the Effectiveness of Sin-Food Taxes

Emilio Gutierrez, Beata Javorick, Wolfgang Keller, Faqiang Li, Ricardo Miranda, Kensuke Teshima, James Tybout

Preliminary and Incomplete
Please Do Not Circulate

this version: November 30, 2023
Please click here for the latest version

Abstract

We study the impact of retail globalization on calorie consumption under alternative policy regimes. Specifically, using household-level surveys and home scanner data, we first examine the effects of Walmart openings in Mexican cities on household consumption patterns. In doing so, we document an eight percent permanent increase in households' purchased calories that coincides with the timing of Walmart openings, and we show that this increase traces to greater consumption of unhealthy foods. Next, we show that when Mexico introduced a tax on highly caloric foods in 2014, caloric intake fell among Walmart shoppers, who substituted for cheaper and healthier food options. Finally, building on Thomassen et al. (American Economic Review, 2017), we estimate a structural model of households' choices concerning the stores they visit and the goods they consume. This model provides a basis for counterfactual analyses of calorie taxes (inter alia), and it allows us to link changes in caloric intake among different types of households to Walmart openings.

1 Introduction

Rising obesity rates and their health consequences are a major concern for both developed and developing nations. Since 1980, obesity rates have almost tripled worldwide (World Health Organization, 2002). One first step toward combating obesity is to recognize that it is caused by increases in the difference between calories consumed and calories burned (Hill, Wyatt, and Peters, 2012). Researchers have proposed two reasons for the increase in calorie consumption: price reductions of high caloric content foods and increases in their availability (Cawley, 2015; Currie et al., 2010; Cutler, Glaeser, and Shapiro, 2003). In response to the increase in obesity rates, sin-food taxes have emerged as a countermeasure. Using the case of Walmart openings and expansions in Mexican cities, we explore how retail globalization, which affects the prices and availability of foods with high caloric content, impacts the composition of household diets and the effectiveness of sin-food taxes.

Mexico offers an ideal setting to investigate the relationship between retail globalization, nutrition, and health policies. It is a country in which the fraction of the population that is obese or overweight has grown the most in recent decades. (According to the Mexican Health and Nutrition Survey, known locally as ENSANUT, in 2012, 35% of the Mexican population was obese), To address this alarming trend, in January

2014, the Mexican government enacted a tax on high-calorie foods and beverages. Moreover, in recent years, Mexico has experienced a rapid increase in the number of Walmart stores, a large foreign retail chain, which has transformed the composition of the retail industry (Atkin, Faber, and Gonzalez-Navarro, 2018; Iacovone et al.,2015). In addition, rich micro-level data are available. We use two types of detailed household data: the National Survey of Household Income and Expenditure, known as ENIGH, a nationally representative household consumption and expenditure survey, and scanner data collected by KANTAR Worldpanel.

In the first part of the analysis, using an event study approach, we document an eight percent permanent increase in the calories purchased by households that coincides with the timing of Walmart openings. We find that this increase is concentrated in unhealthy foods. Moreover, we find that after entry, households source the largest part of their weekly consumption of packaged products from Walmart stores more frequently. In the second part of the analysis, we study the effectiveness of the tax on high-caloric content foods imposed by the Mexican government in January 2014. We find that consumers who frequently source most of their packaged product consumption from Walmart reduced their total caloric intake as opposed to the rest of consumers. The reduction of calorie consumption from untaxed products is explained by the fact that Walmart's advantage is smallest at the middle range of calorie intensity, which induced consumers to substitute cheaper and healthier products. We rationalize these results with a grocery shopping model developed by Thomassen et al. (2017).

Our study is related to several strands of the literature. First, it relates to studies linking globalization and obesity. Here, cross-country studies find mixed results. For example, Miljkovic et al. (2015) and Vogli et al. (2014) find a positive and significant association between trade openness and obesity and body mass index, while no such association is found in Oberlander, Disdier, and Etilé (2017) or Costa-Font and Mas (2016). Giuntella, Rieger, and Rotunno (2017) analyze food imports and obesity in Mexico using state-level data. Our study is unique in focusing on the role of retail globalization in both rising obesity and the effectiveness of policies against it.

Our study also contributes to the literature on food deserts and the nutritional implications of shopping costs. While we do not find that Walmart entries on their own improve the diet of households, we show how retail globalization increases the availability of alternatives to high caloric content foods in developing countries (Rose and Richards 2004; Morland et al., 2002 ; Cotterill and Franklin, 1995 ; Weinberg, 1995) and we document how this allowed consumers to substitute toward a wider variety of products and sizes after the tax enactment (Allcott et al., (2019)[1]; Chung and Myers Jr, 1999; Kaufman, 1998; Kaufman et al., 1997). This result, along with our finding that after entry, households start sourcing most of their packaged products consumption from Walmart stores, points in the direction of studies that have analyzed the implications of travel costs and one-stop shopping, such as Thomassen et al. (2017). The implications of these purchasing behaviors for the effectiveness of sin taxes remain to be studied.

[1]While Allcott et al., (2019) do not find that supermarket entries affect nutritional outcomes, our setting differs from theirs in at least two aspects. First, travel costs are different in Mexico than in the USA. For instance, only 50% of the households in our sample have a car. Second, the alternatives to globalized retail that domestic retailers offer differ between developed and developing countries.

Our study is also related to the growing literature that has linked Walmart and other supermarket openings to the body mass index and diet composition of US households. (Volpe, Okrent, and Leibtag, 2013; Courtemanche and Carden, 2011) [2] Our contribution to this literature is to document whether the relationship between Walmart openings and the composition of the diets of households is different in Mexico than in developed countries.

Our paper also contributes to the understanding of the impact of retail globalization in low and middle-income countries (Iacovone et al. 2015; Javorcik and Li 2013; Javorcik, Keller, and Tybout, 2006;). While some existing papers have focused on Mexico and Walmart openings, they differ from ours in terms of methodology and research questions. In particular, our study is the first to investigate the impact of Walmart on the caloric intake and diet composition of Mexican households. The closest to our study in terms of methodology and context is Atkin, Faber, and Gonzalez-Navarro (2018). They use an event-study methodology (using foreign supermarket entries as the event of interest) for the Mexican context. However, their analysis is broader, as it measures the impact of supermarket stores' entry on overall prices and welfare (and its distribution).

Finally, this work is linked to papers that have explored the relationship between food prices, taxes on foods and beverages and the diet of households, or obesity rates in Mexico (Aguilar, Gutierrez, and Seira, 2019; Colchero et al., 2016; Gracner, 2015) and other contexts (Harding and Lovenheim, 2017; Dubois, Griffith, and Nevo, 2014; Grossman, Tekin, and Wada, 2014; Fletcher, Frisvold, and Tefft, 2010; Powell and Chaloupka, 2009; Beydoun, Powell, and Wang, 2008). The contribution of our work to this literature is to identify retail globalization as one of the forces behind these price changes and analyze its implications.

The paper is organized as follows. The next section describes the data sources exploited in the paper. Section 3 presents the empirical strategies employed in all the analyses presented. Section 4 presents the results. Section 5 introduces a quantitative model of grocery shopping that helps explain the key data patterns. Section 6 introduces the procedure of structural estimation.

2 Data

2.1 ENIGH: The composition of households' diets

To observe the composition of households' diets following the entry of Walmart to Mexican municipalities, we use the ENIGH surveys administered every two years. Although the municipalities and households covered change from one survey to another, they are designed to be representative at the state level. For this reason, all surveys include households from most of the large municipalities of Mexico. For our analysis, we use the ENIGH surveys from 2008, 2010, 2012, 2014, and 2016.

The main goal of the ENIGH surveys is to capture expenditures and incomes. Households are asked to report all their food purchases during the three months prior to the survey. Expenditures are classified into multiple categories, one of which corresponds to foods and beverages. This category is partitioned into 247

[2] Holmes (2011); Jia (2008); Hausman and Leibtag (2007) and Basker (2005) explore the relationship between Walmart store openings and other outcomes for the US context.

different subcategories, 211 of which correspond to foods, 24 to beverages, and the remaining 12 to tobacco, food for animals, meals eaten outside the household, and in-kind transfers.

While ENIGH includes detailed information on expenditures and quantities purchased for each of the subcategories listed, it does not contain nutritional information. To assign caloric contents to each of the foods and beverages subcategories, we use the National Nutrient Database for Standard Reference (NNDSR) published by the United States Department of Agriculture.

The NNDSR is a data set on the nutritional content of most products consumed in the United States. It comprises 1,137 categories, and, in total, 85% of the food subcategories of the ENIGHs are covered. We collected caloric content information on the ENIGH subcategories not included in the NNDSR, mainly traditional Mexican foods. When the NNDSR lists products at a more detailed level than ENIGH, we assign the average calories per kg/liter of NNDSR categories to the corresponding unique subcategory in ENIGH.

Apart from expenditures, the ENIGH surveys include detailed information about the sociodemographic characteristics of households. They include household size, living-place characteristics, and information on all household members' age, employment, and health. The exact location of households is not provided, but the surveys report the population of the localities in which households are located within municipalities.

2.2 KANTAR: Purchases of packaged and taxed products.

In addition to the ENIGH surveys, we exploit scanner panel data on households' consumption of packaged products collected by KANTAR Worldpanel. This is a high-frequency data set that registers the purchases of households at the store and bar code levels. In addition to stores and bar codes, it registers the quantities and prices involved in each purchase and the exact date at which each transaction occurred. Our data cover the period from 2011 to 2015.

As in the case of the ENIGH surveys, the KANTAR data set does not include information on the nutritional content of bar codes. Thus, we obtain it from a database on the nutritional content of packaged products in Mexico that was specifically collected to be merged with the KANTAR Worldpanel data by Aguilar, Gutierrez, and Seira (2019). Nutritional content was directly collected for 71% of the bar codes in the data set. This represents 68% of the observed expenditures. For the remaining 29% of bar codes, caloric content was imputed from those for which caloric content was directly collected. The nutritional imputation is made at the bar code level.

2.3 Walmart store entries

Our main data source on the dates at which Walmart stores entered the Mexican market is Walmart's monthly financial reports. These can be found on the Walmart of Mexico website. They indicate the exact month and city of store openings. From the financial reports, we collected the month in which 55 Walmart supercenters opened in the cities covered by the KANTAR data set and the opening months of an additional 30 stores in the cities covered by the ENIGH surveys. The entries occurred from 2012 to 2015.

Because the ENIGH surveys are collected every two years, most of the entries we determine from the Walmart bulletins occurred between ENIGH 2012 and ENIGH 2014. This limits the variation in the entries' timing required to perform an event study analysis. To address this issue, we complement our information on entries using the registries of Walmart supercenters in the Mexican National Directory of Economic Units (DENUE). This registry is updated yearly for economic units with more than 100 employees. The frequency of these updates is sufficient for us to discern which Walmart store entries occurred in the period between ENIGH 2010 and ENIGH 2012, which increases the time variation of the entries in our sample. From this exercise, we recover 44 additional entries outside the time covered by the financial reports of Walmart of Mexico.

3 Empirical strategy

In this section, we discuss our method to estimate the effect of Walmart entries on households' calorie consumption and purchasing patterns. In addition, we explain how we test whether consumers respond differently to a tax on high caloric content (HCC) foods based on whether they buy at Walmart or other stores.

3.1 Effect of Walmart store entries on calorie consumption

We use an event study approach based on the ENIGH surveys to estimate the effect of Walmart entries to Mexican cities. These data sets cover all household food purchases, enabling us to observe the whole composition of households' diets and how they changed in response to the Walmart entries.

We include all entries from Walmart of Mexico bulletins between 2012 and 2015 in our regressions. As noted above, to expand the period covered by our analysis and achieve sufficient time variation to perform an event study, we use DENUE as an additional source of entries. DENUE is updated yearly for stores with over 100 employees, which is the case for Walmart supercenters. ENIGH surveys are collected every two years. Hence, for an event study, the registry dates to DENUE are sufficient to define pre- and post-entry surveys.

In the ENIGH surveys, municipalities are further divided into localities. From the DENUE data, we know that 93% of Walmart supercenters are in localities with more than 15,000 inhabitants. We restrict the sample for our main analysis to those localities.[3] In our analysis, we retain only those municipalities appearing in at least two surveys before and two after Walmart entries. We repeat those observations when more than one entry occurs in the same city but on different dates relative to the ENIGH surveys. This leaves 93 different entry time–city combinations, representing 129 Walmart store openings distributed across 80 municipalities over four years.

We are interested in the caloric intake of households and the product types in which the changes in diets

[3] Results for localities with less than 15,00 inhabitants are excluded from this writing sample. We find no effect of Walmart store entries on households' diets in those localities.

are concentrated. Following Hut and Oster (2015), we classify all the products that appear in the ENIGH surveys as healthy if they are "obviously healthy," *i.e.*, products that unambiguously are not harmful to human health. The healthy category includes fruits, vegetables, low-fat, and fresh sources of protein, such as fish and chicken. It excludes packaged products, such as cereals, candies, snacks, sodas, processed juices, and prepared meals. It also excludes foods with high fat, such as pork and beef. All the excluded products are labeled as unhealthy. To measure caloric intake, we exploit information from the NNDSR and compute the monthly caloric intake of households from healthy products, unhealthy products, and the sum of calories from both categories.

As noted above, the households in the ENIGH change from one survey to another; therefore, our panel is balanced at the municipality level but not at the household level. For this reason, we control for households' observables in addition to time/survey and municipality-fixed effects. Purchased calories depend on households' size; households' tastes might vary across cities. Therefore, we include interactions between city indicators and household size. Moreover, consumption is also likely to vary at different life stages. For this reason, we control for household head age interacted with time fixed effects in our regressions. Our event study specification is as follows:

$$log\ Y_{tmh} = \sum_{k=-2}^{3} \delta_k \mathbf{I}_{\{t-E_{jm}=k\}} + \tau_t + \eta_m + \alpha_t a_h \mathbf{I}_{\{t\}} + \beta_m s_h \mathbf{I}_{\{m\}} + \varepsilon_{tmh} \tag{1}$$

Where Y_{tmh} is calories from the category of interest consumed by household h at time t. The function $\mathbf{I}_{\{t-E_{jm}=k\}}$ indicates the ENIGH rounds relative to entry. τ_t and η_m are time and municipality fixed effects, respectively. a_h denotes household head age and s_h denotes household size. The functions $\mathbf{I}_{\{t\}}$ and $\mathbf{I}_{\{m\}}$ are time and municipality indicators. Finally, ε_{tmh} is an error term.

3.2 Timing, purchasing patterns and calories from packaged products

Using the ENIGH surveys, we can assess the long-term effects of Walmart entries on households' diets. However, we do not observe the exact timing of the consumption changes or the stores where the purchases occur. We repeat our event study analysis using the KANTAR scanner panel data to investigate these issues. Here, we can observe monthly consumption and the stores where purchases occur.

To ensure that our household panel is balanced, we restrict our sample to households for which we can observe consumption for at least six months before and after a Walmart entry. As in our analysis of the ENIGH surveys, we repeat observations when multiple entries to the same city occur at different times. This leaves us with 2,576 households in 28 cities, for which we observe 55 Walmart entries. Because we want to observe the consumption of households at the highest possible frequency (monthly), we only use entries from the Walmart of Mexico bulletins.

Having household panel data allows us to include household fixed effects in our analyses and control for unobserved tastes and characteristics that affect calorie consumption and purchasing patterns. Thus, our event study is performed at the household–month level. Our specification is as follows:

$$lY_{tmh} = \delta_{<-6}\mathbf{I}_{\{t-E_j=<-6\}} + \sum_{k=-6}^{6} \delta_k \mathbf{I}_{\{t-E_j=k\}} + \delta_{>6}\mathbf{I}_{\{t-E_j=>6\}} + \tau_t + \eta_h + \beta_m t \mathbf{I}_{\{m\}} + \varepsilon_{tmh}. \tag{2}$$

Where Y_{tmh} is the outcome of interest for household h's at time t in city m. $\mathbf{I}_{\{t-E_j=k\}}$ indicates months relative to entry, τ_t and η_h are time and household fixed effects, respectively, and ε_{tmh} is an error term. $\mathbf{I}_{\{m\}}$ is a city indicator; we include city time trends represented by $\beta_m t$.

Our main outcome variable is the total calories each household purchases. We are also interested in observing how purchasing patterns change after entry. In particular, we want to know if there is an increase in purchases at Walmart and if Walmart becomes the main store from which households source their consumption of packaged products. This serves two purposes: first, it confirms that our entry dates actually coincide with store openings, and second, it helps us understand the relevance of Walmart as a retailer for households' shopping decisions. Hence, for every month and every household, we compute the proportion of weeks in which a Walmart store had the largest share of the weekly observed expenditures. Therefore, it was the household's main source of packaged products. Finally, we measure the share of total expenditures and calories corresponding to Walmart purchases during weeks in which a Walmart store was the households' main source of packaged products.

3.3 Effect of the tax on high caloric content foods on calorie consumption and substitution

It is possible that the prices and availability of products might differ between Walmart and other retailers. This may induce differences in consumers' responses to price changes depending on where they buy groceries. In particular, the response to changes in the price of HCC products due to the introduction of the tax on HCC foods might differ between Walmart customers and customers of other stores. We explain how we test this hypothesis.

First, we classify all products that appear in the KANTAR data set. We begin by identifying the products that are subject to the tax. The tax applies to all beverages that have added sugar and to all foods with more than 275 kilocalories per 100 grams that are not considered an essential component of the diet of Mexicans (such as oil or tortillas). We define all taxed products as HCC products. Then, we divide untaxed bar codes into two broad categories: foods and beverages. Within each category, products that fall in the first quartile of total calories are defined as low caloric content (LCC) products, and all remaining products are defined as middle caloric content (MCC). The definition of LCC products does not imply that these products are healthier because both caloric density and size must be considered. For instance, sufficiently small presentations of unhealthy products could be labeled as LCC products.

To test which products are relatively cheaper at Walmart, we exploit the KANTAR data set at its most disaggregated level. We regress the log of observed purchase prices on dummies that indicate if a purchase occurred at Walmart and its interactions with indicators of the type of product that was bought. We are interested in the sign of the Walmart indicator and its interactions. Our specification is as follows:

$$log\ P_{bsmp} = \gamma_b + \mu_m + \omega \mathbf{I}_{\{s=w\}} + \sum_{i \in h,m,l} \beta_{iw} \mathbf{I}_{\{s=w\}} \mathbf{I}_{\{b \in i\}} + \varepsilon_{bsmp}, \tag{3}$$

where P_{bsm} is the price of bar code b at store s in month m for purchase p. γ_b is a barcode fixed effect and μ_m is a month fixed effect. $\mathbf{I}_{\{s=w\}}$ is an indicator of purchases occurring at Walmart, and $\mathbf{I}_{\{b \in i\}}$ is an indicator of product type (HCC, MCC or LCC). ε_{bsmp} an error term.

Furthermore, we must examine how the composition of households' diets changes due to the tax and if this effect is the same for Walmart customers as for the rest of consumers. To do this, we classify households as Walmart customers or not according to their pre-tax purchasing behavior. We proceed as follows: first, we identify for each household the weeks in which the largest share of their expenditures in packaged products was spent at a store belonging to Walmart. Then, we define Walmart customers as all the households that used Walmart as their main store for at least one week of the month for at least nine out of the twelve months in 2013. While these variables are defined with respect to weekly expenditures, we aggregate them by month so that the level of measurement coincides with the time unit at which we observe entries and defines the frequency of our panel.

To analyze purchases of packaged products, we compute the total amount of purchased calories from all product types according to the definition used to analyze prices (HCC, MCC, and LCC products) and regress them on an indicator of the tax enactment and its interaction with a variable that indicates Walmart customers. We repeat this analysis for the total calories purchased by households, for calories from taxed and untaxed products, and, finally, separate the calories from untaxed products into calories from LCC and MCC products. Our regression equation is:

$$log\ Y_{ht} = \eta_h + \tau_t + \beta_T \mathbf{I}_{\{t \in T\}} + \beta_{TW} \mathbf{I}_{\{t \in T\}} \mathbf{I}_{\{h \in W\}} + \varepsilon_{ht} \tag{4}$$

Where Y_{ht} represents total purchased calories by household h from the category of interest (HCC, MCC, LCC, and their relevant combinations) at time t. Time and household fixed effects are represented by η_h and τ_t. The tax enactment is indicated by $\mathbf{I}_{\{t \in T\}}$ and $\mathbf{I}_{\{h \in W\}}$ indicates that household h is a Walmart customer according to its pre-tax purchasing behavior. The error term is ε_{ht}.

Note that if there is a positive association between Walmart entries and caloric intake, households who start attending Walmart after the tax would experience an increase in their caloric intake. This could be a confounding factor for our analysis. Therefore, we exclude from our sample households who had no purchases at Walmart during 2013 but started attending it in 2014 or 2015.

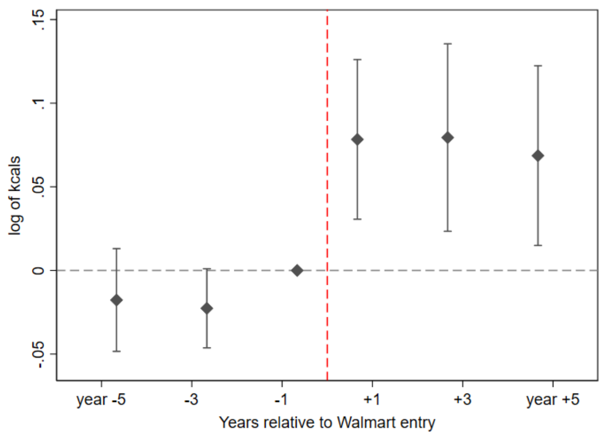
4 Empirical Results

In this section, we discuss the results of our event study. We find that there was an increase in the calorie consumption of households after the Walmart entries. In addition, we provide evidence that, after the enactment of the tax, households' substitution of taxed products with untaxed products differed depending on whether they shopped at Walmart or other stores. We conclude that the tax was more effective in reducing calorie consumption for households that could buy from Walmart.

4.1 Effect of Walmart store entries on calorie consumption

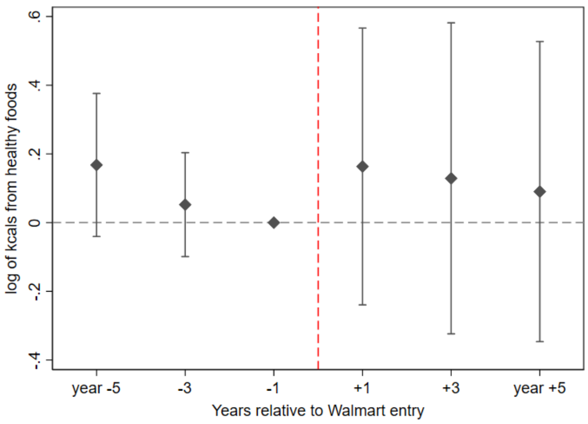
Our event study estimations using the ENIGH surveys show that Walmart's entry transformed households' diets into Mexican municipalities. Figure 1 shows the event study coefficients for the ENIGH surveys before and after the Walmart entries. We find that there was a permanent 8% increase in households' total calorie consumption after Walmart entered the Mexican market. Still, there is no evidence of an increasing trend before its entry. Figures 2 and 3 disaggregate the effect on healthy and unhealthy products. We find that the effect comes from a 6% increase in calories from unhealthy products. Our findings are summarized in columns 1 to 3 of Table 1, where we report the average effect for all post-entry periods on total caloric intake and calories from healthy and unhealthy products.

Figure 1: Effect of Walmart entries on calorie consumption of households.



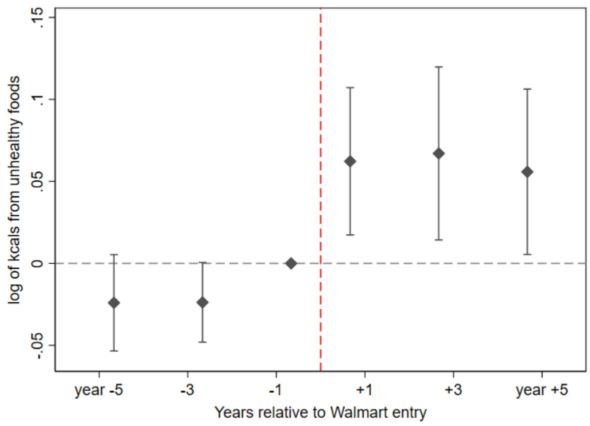
Notes: This figure depicts the event study coefficients for the log of households' calorie consumption from healthy products. The coefficients correspond to each ENIGH survey immediately before and after Walmart entries. The sample is restricted to municipalities that appear in at least two surveys before and after Walmart entries.

Figure 2: Effect of Walmart entries on households' calorie consumption from healthy products.



Notes: This figure depicts the event study coefficients for a log of households' calorie consumption from healthy products. The coefficients correspond to each ENIGH survey immediately before and after Walmart store entries. The sample is restricted to municipalities that appear at least in two surveys before and after Walmart entries.

Figure 3: Effect of Walmart entries on households' calorie consumption from unhealthy products.



Notes: This figure depicts the event study coefficients for the log of households' calorie consumption from unhealthy products. The coefficients correspond to each ENIGH survey immediately before and after Walmart store entries. The sample is restricted to municipalities that appear at least in two surveys before and after Walmart entries.

Table 1: Event study estimation of the effect of Walmart entries on Households calorie consumption

ENIGHs			KANTAR			
(1) log of calories All products	(3) log of calories Healthy products	(3) log of calories Unhealthy products	(4) log of purchased calories	(5) Prop. of weeks during which WM was the main store	(6) Share of exp. From WM as main store	(7) Share of cals. from WM as main store
-0.025	0.0311	-0.025	0.00	0.001	0.00261	0.002
(0.015)	(0.088)	(0.015)	(0.007)	(0.001)	(0.001)	(0.001)
0.079***	0.164	0.063**	0.014*	0.006***	0.007***	0.005***
(0.029)	(0.246)	(0.027)	(0.007)	(0.001)	(0.002)	(0.001)
Yes	Yes	Yes				
Yes	Yes	Yes				
Yes	Yes	Yes				
Yes	Yes	Yes				
			Yes	Yes	Yes	Yes
			Yes	Yes	Yes	Yes
0.226	0.104	0.262	0.576	0.542	0.547	0.533
232,795	232,795	232,795	381,177	381,177	381,177	381,177

Notes: Estimates for the effect of Walmart entries on the calorie consumption of households. Results are shown aggregating across all pre- and post-entry periods for four years in the case of ENIGH surveys and six months in the case of the KANTAR data set. For ENIGH, the sample is restricted to households in municipalities that appear at least in the two consecutive periods before and after entries. Results are reported in columns 1 to 3. The first column shows results for all calories in products purchased by households. Columns 2 and 3 show results for calories from healthy and unhealthy products. For the KANTAR data set, regressions are performed using a household monthly scanner panel perfectly balanced for the six months before and after entry. Results are shown in columns 4 to 7. Column 4 reports results for the log of purchased calories. For columns 5 to 7, we identified the weeks in which a single Walmart store had the largest share of total observed expenditures and, therefore, was the main store from which packaged products were purchased. Column 5 reports the effect of entries on the proportion of weeks in a month during which a Walmart store was the households' main source of packaged products. Columns 6 and 7 report the effect of entries on the share of total expenditures and purchased calories corresponding to the products bought from Walmart stores when they were the households' main source of packaged products.

4.2 Timing, purchasing patterns, and calories from bar-coded products

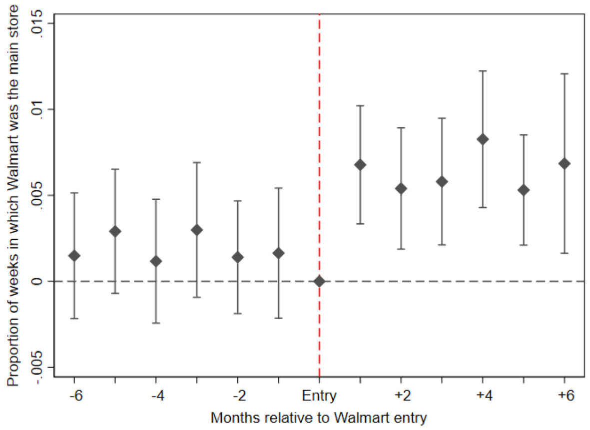
Because of the low frequency of the ENIGH surveys, our estimates in the previous section could raise the question of whether the increase in calorie consumption coincides with the Walmart store entries. We address this issue using the KANTAR household scanner panel data and find that calorie increases from bar-coded products coincide with the months in which Walmart stores enter the Mexican municipalities. Figures 4 to 6 depict the event study coefficients for the six months before and after Walmart stores entries based on the KANTAR data. We find a 1.4% increase in calories from purchased packaged products. Although the KANTAR results show the effect of entries on only a portion of households' consumption, our findings are consistent with those from the ENIGH surveys. Hence, we conclude that Walmart stores' entries are behind Mexicans' transition toward more calorie-intense diets.

Moreover, recently opened Walmart stores have become one of the main sources of packaged products for households. Figure 4 shows a significant increase of 0.6 percent in the proportion of Weeks per month during which a Walmart store had the largest share of observed households' expenditures. Moreover, the share of total expenditures corresponding to purchases at Walmart stores during weeks where the households' main source of packaged products significantly rises by 0.7 percent. We observe a similar pattern for the share of calories from Walmart as the main store that increases by 0.05 percent as depicted in figure 6.

Table 1 summarizes our findings for the effect of Walmart entries based on aggregating the estimation results for the six months before and after entry. We find no evidence of pre-entry trends in households' purchased calories or the frequency with which Walmart stores are the main source of packaged products for households in our sample.

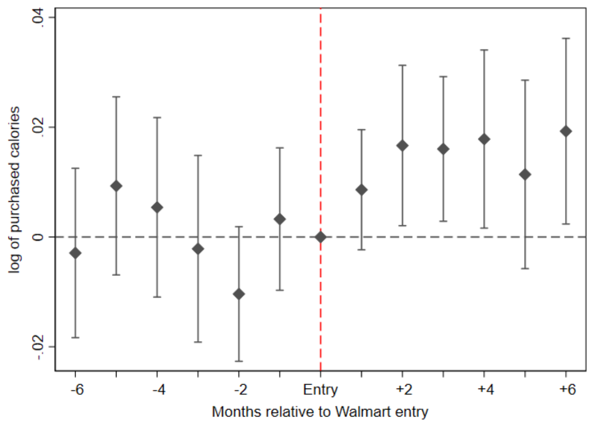
Note that the identified effects correspond to the average change among households in the city of entry. While, on average, there is a significant increase in the count of weeks in which Walmart is the main store from which packaged products are bought, this effect is unlikely to be homogeneous across households. Only 200 out of 2576 households in our data set used Walmart as their main store for at least one week every month after entry, and only 18 households used Walmart as their main store every week after entry. Walmart becomes a relevant retailer after entry, but other retailers remain the primary source of packaged products for many households.

Figure 4: Entries and Walmart as main source of packaged products



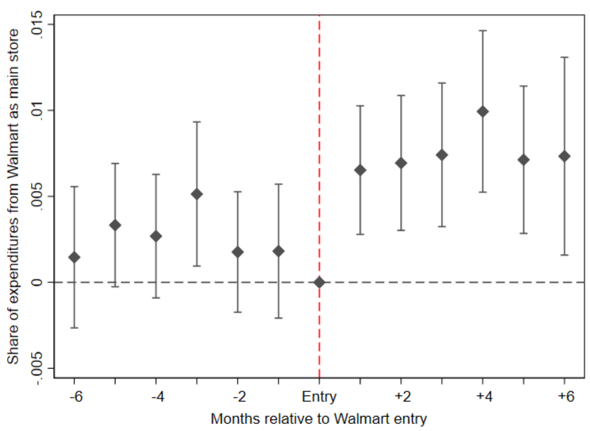
Notes: This figure depicts the event study coefficients for the proportion of weeks per month in which a single Walmart store had the largest share of total observed expenditures and, therefore, was the household's main source of packaged products. The coefficients correspond to each of the six months before and after entry. The sample is a perfectly balanced panel of households. The sample is restricted to cities that experience at least one entry during the sampling period.

Figure 5: Effect of Walmart entries on households' purchased calories from packaged products.



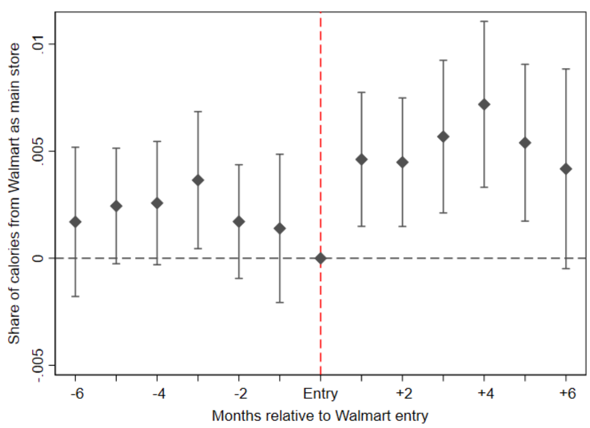
Notes: This figure depicts the event study coefficients for the log of households' observed purchased calories in the KANTAR data set. The coefficients correspond to each of the six months before and after entry. The sample is a perfectly balanced panel of households. The sample is restricted to cities that experience at least one entry during the sampling period.

Figure 6: Entries and the share of total expenditures from Walmart as the main source of packaged products.



Notes: For this figure, we identified the weeks in which a single Walmart store had the largest share of total household observed weekly expenditures. Then, the dependent variable is the share of monthly expenditures in the identified store-week combinations (*i.e.* when a Walmart store was the households' main source of packaged products). The figure depicts the event study coefficients for each of the six months before and after entry. The sample is restricted to cities that experience at least one entry during the sampling period.

Figure 7: Entries and the share of total calories from Walmart as the main source of packaged products.



Notes: For this figure, we identified the weeks in which a single Walmart store had the largest share of total household observed weekly expenditures. Then, the dependent variable is the share of monthly purchased calories from products purchased at the identified store-week combinations (*i.e.* when a Walmart store was the households' main source of packaged products). The figure depicts the event study coefficients for each of the six months before and after entry. The sample is a perfectly balanced panel of households. The sample is restricted to cities that experience at least one entry during the sampling period.

4.3 Effect of the tax on calorie consumption and substitution

The second goal of this work is to determine if the change in the composition of the diets of households induced by the tax on HCC foods was different for those who bought from Walmart than for other households. Moreover, we want to understand the forces that could drive this differential response.

Our analysis of prices shows that, on average, all products are 6% cheaper at Walmart than at other stores. Moreover, we find that both taxed and LCC packaged products are relatively cheaper at Walmart compared with other stores. These results are summarized in Table 2. These differences in prices and relative prices suggest the mechanism by which the response to the tax might vary between consumers depending on where they shop.

Second, our estimates for the responses of households to the tax enactment show that substitution patterns were different depending on whether they shopped at Walmart or elsewhere. Column 1 of Table 3 shows that by Aguilar, Gutierrez, and Seira (2019), the tax had no significant impact on the total calorie consumption from bar-coded products of the households in our sample. However, we find that the tax caused an effective reduction in the total amount of calories purchased by Walmart customers. This is explained by the fact that the increase in calories from purchases of untaxed products induced by the tax was significantly lower for Walmart customers than other consumers. When disaggregating this effect, we find that Walmart customers substituted significantly more LCC products, and their consumption of MCC products increased less than for the rest of households. Hence, we conclude that the tax on HCC foods was more effective in reshaping the composition of household diets for Walmart customers than it was for other consumers.

Table 2: Prices in Walmart by type of product relative to the rest of stores during 2013 and 2014.

	(1)	(2)	(3)	(4)	(5)	(6)
	Pre tax		Post tax		All periods	
Variables	log Price	log Price	log Price	log Price	log Price	log Price
Walmart	-0.0631***	-0.0396***	-0.0665***	-0.0488***	-0.0660***	-0.0452***
	(0.00426)	(0.00286)	(0.00274)	(0.00282)	(0.00404)	(0.00306)
Walmart#Taxed (HCC)		-0.0488***		-0.0380***		-0.0442***
		(0.00414)		(0.00293)		(0.00377)
Walmart#Untaxed Low caloric cont. (LCC)		-0.0173***		-0.0135***		-0.0134***
		(0.00440)		(0.00392)		(0.00322)
Barcode FE	Yes	Yes	Yes	Yes	Yes	Yes
Month of year FE	Yes	Yes	Yes	Yes	Yes	Yes
Period	2013	2013	2014	2014	2013-2014	2013-2014
R-squared	0.953	0.953	0.948	0.949	0.944	0.944
Observations	6367353	6367353	6227351	6227351	12594704	12594704

Robust standard errors clustered at the store level in parentheses
*** p<0.01 ** p<0.05 * p<0.1

Notes: The table shows estimation results for prices in Walmart compared to other stores. The dependent variable is the log of the observed purchase price. The observations correspond to all the registered transactions at the date–store–bar-code level made by households in the KANTAR data set from 2013 to 2014. Each observation is weighted by the quantities purchased. Walmart is a dummy that indicates the purchase occurring at Walmart. Taxed indicates all purchases of beverages and foods subject to the tax. Untaxed Low caloric cont. (LCC) indicates beverages and foods in the lowest quartile of caloric content among products not subject to the tax. Untaxed products with caloric content above the first quartile are the excluded category. The first two columns were estimated using only pre-tax observations, whereas columns 3 and 4 were estimated using only post-tax observations. The last two columns were estimated using all observations.

Table 3: Walmart and substitution. Observed effect of the tax and the option to attend Walmart on the calorie consumption of households.

Variables	(1)	(2)	(3)	(4)	(5)
	Purchased calories				
	All products	Taxed products	Untaxed products		
	All	HCC	MCC and LCC	MCC	LCC
Tax	227.5	-1,503***	1,730***	1,698***	31.80
	(340.8)	(187.8)	(215.8)	(208.6)	(55.71)
Tax#Walmart customer	-1,599**	-536.5	-1,063**	-1,278***	215.2**
	(702.3)	(353.0)	(469.5)	(450.5)	(101.7)
Household FE	Yes	Yes	Yes	Yes	Yes
Month of year FE	Yes	Yes	Yes	Yes	Yes
Dep. var mean	78,872	34,479	44,393	39,549	4,843
Tax+Tax#Walmart F-statistic	4.989	46.56	2.562	1.109	8.427
p-value	0.0256	0	0.110	0.292	0.00372
R-squared	0.540	0.559	0.477	0.482	0.484
Observations	131,796	131,796	131,796	131,796	131,796

Notes: The table shows the effect of the tax enactment on the calorie consumption of Walmart customers and the rest of households in the KANTAR data set. Walmart customers are defined according to their pre-tax purchasing behavior using the following procedure: For each household, we identify if there were weeks in which a Walmart store had the largest share of observed expenditures among all the stores from which the household bought at least one item. Then, a household is defined as a Walmart customer if this was the case for at least one week of the month for at least nine out of the twelve months of the year. To avoid the bias induced by households who started attending Walmart in 2013 and therefore increased their calorie consumption, all households who made no purchases at Walmart during 2013 but began shopping at Walmart in 2014 or 2015 were excluded from the sample. The first column shows the effect on purchased calories from all products. Columns 2 and 3 show the effect on purchased calories from taxed (High caloric content products, HCC) and untaxed products, respectively. The last two columns disaggregate the effect shown in column 4 into MCC untaxed products (i.e., untaxed product bar codes associated with caloric contents above the first quartile among the untaxed products) and LCC untaxed products (untaxed product bar codes with caloric content within the first quartile of untaxed products).

5 Model

This section introduces a quantitative model characterizing Mexican household's demand for groceries. We use the model to examine the effects on household grocery choices and calorie consumption of Walmart openings under different tax regimes on sin-food.

We use the model developed by Thomassen et al. (2017) to characterize Mexican households' shopping patterns. Heterogeneous consumers choose the products they purchase, the stores they will shop at, and the number of trips they will make. Different stores carry different bundles of products, and product prices are store-specific. Further, stores differ in terms of the commute time required to reach them, which in turn varies across households. So, taking stock of their own tastes and budgets, households must weigh commute times, product offerings, and prices when choosing which store or stores to visit.

This model captures . . . We have shown that Walmart typically offers a broader range of products in terms of variety across and within categories. Moreover, different stores could offer different varieties of products or products with distinct characteristics. Therefore, we need a model to capture the cross-category substitution of households and allow it to influence the substitution across stores to comprehensively quantify the effects of Walmart's openings and the sin-food tax. For example, when the price of calorie-intensive food in stores owned by Walmart goes up because of the introduction of the sin-food tax, households who chose those stores may switch to more healthy food from the same stores or different stores.

In the following, we start with a household's weekly purchase problem and characterize the composition of the household's utility and decisions. Subsequently, we derive the solution to the decision problem. Lastly, we point out the source of heterogeneity of households' tastes in the model.

5.1 Utility

The model characterizes a household's weekly problem in grocery purchasing. In a week, the household chooses how much to purchase in each category of grocery and which stores to visit for each category. Denote each available store of the household by j and each category by k. Denote all the available stores as $\mathcal{J}$ and denote the total number of stores as $J = |\mathcal{J}|$. Denote the number of all categories as K. To make the model tractable, we assume that the household visits no more than two stores in a week and uses only one store per purchase category. We provide evidence that supports these assumptions in the empirical strategy section.

We refer to shopping choice to be the set of stores the household visited, denoted by c. If store j and j' are chosen, $c = \{j, j'\}$. If only store j is chosen, $c = \{j\}$. Denote the set of all the possible shopping choices as $\mathcal{C}$. According to its definition, $\mathcal{C}$ contains all the possible pairs of stores from $\mathcal{J}$ as well as all the singleton (as a set) in $\mathcal{J}$. A shopping cost $\Gamma(c)$ is associated with each choice of c. This cost depends on the number of stores in c, $n(c)$, and the sum of distances traveled to each store in c. It can be considered the traveling expenses and time wasted on the trip. We refer to grocery choice, denoted by q_{jk} for each category k and store j as the nonnegative continuous quantities the household purchases for category k in store j. To facilitate exposition, we use two vectors, $\mathbf{q}$ and $\mathbf{d}$, to fully characterize all q_{jk}s decided by the household. The $\mathbf{q}$ is a $K \times 1$ vector characterizing the quantity that the household purchases for all K categories. The $\mathbf{d}$ is the stores chosen for all categories from a set of chosen stores. As an illustration, suppose $K = 3, c = \{\text{Walmart}, \text{Soriana}\}, \mathbf{q}, = (1.23, 0.19, 0), \mathbf{d} = (\text{Walmart, Soriana, Soriana})$. This means that the household goes to Soriana and Walmart and buys 1.23 units of the first category in Walmart, 0.19 units of the second grocery, and zero in the third grocery in Soriana. We use $\mathcal{D}_c$ to indicate the set of possible alternatives for $\mathbf{d}$ conditional on the shopping choice c. Let $\mathbf{p}$ be the full vector of store-category prices p_{jk} and μ the full vector of store-category tastes μ_{jk}.

Then, the household's utility under the decision $(c, \mathbf{d}, \mathbf{q})$ at prices $\mathbf{p}$ is given by

$$\begin{aligned} U(c, \mathbf{d}, \mathbf{q}, \mathbf{p}) &= u(\mathbf{q}, \mathbf{d}) - \alpha \mathbf{p}_{\mathbf{d}}' \mathbf{q} - \Gamma(c) + \varepsilon_c \\ &= (\boldsymbol{\mu}_{\mathbf{d}} - \alpha \mathbf{p}_{\mathbf{d}})' \, \mathbf{q} - 0.5 \mathbf{q}' \Lambda \mathbf{q} - \Gamma(c) + \varepsilon_c. \end{aligned} \tag{5}$$

From equation (5), the household's utility consists of four parts: $u(\mathbf{q}, \mathbf{d}) = \mu_{\mathbf{d}}' \mathbf{q} - 0.5 \mathbf{q}' \boldsymbol{\Lambda} \mathbf{q}$ is the utility from the weekly consumption; $-\alpha\mathbf{p}_{\mathbf{d}}'\mathbf{q}$ is the distutility from the expenditure on groceries and α measures the price sensitivity; $-\Gamma(c)$ is the shopping cost; and ε_c is the household's idiosyncratic taste shock. We use $\mu_{\mathbf{d}}$ and $\mathbf{p}_{\mathbf{d}}$ to indicate the tastes and prices (respectively) for all categories given the store-category choice $\mathbf{d}$. The $\mathbf{\Lambda}$ is a symmetric $K \times K$ matrix of parameters, assumed to be common across consumers.

As is emphasized in Thomassen et al. (2017), this model has two sources of product differentiation: between stores at the category level and across shopping choices c at the level of fixed utility, captured in $\Gamma(c)$. The first source of differentiation stems from the fact that the household views stores differently for any category (captured in the store-category taste vector μ). This source of differentiation is critical in Mexico's context because considerable heterogeneity exists in terms of the availability of certain stores across households. Moreover, different stores offer different varieties of groceries, and the difference in availability of stores further implies the heterogeneity in the offering of categories and their prices. For example, Walmart offers a broader range of groceries, and its products' prices and calorie content are different and influenced differently by the new tax. Additionally, the availability of a particular store can vary within a household over time, even if the household does not relocate. This is because of the entry of chain grocery stores like Walmart during our sample period, as emphasized in the previous sections. The second source of product differentiation comes from the spatial variation in store locations relative to the consumer, which is associated with different shopping costs across households. This heterogeneity is also influenced by the entry of chained grocery stores. To reflect all such heterogeneity, we allow households to be heterogeneous, and each of them can be fully characterized by a trio of parameters: $(\mu, \alpha, \mathbf{\Gamma})$ up to the taste shock ε_c which is assumed to distribute as an i.i.d. type-I extreme value distribution across shopping choices of the household. We introduce the functional form of the $(\mu, \alpha, \mathbf{\Gamma})$ in Section 5.3.

5.2 Demand

The household maximizes $U(c, \mathbf{d}, \mathbf{q}, \mathbf{p})$ by choosing $(c, \mathbf{d}, \mathbf{q})$:

$$\max_{c \in \mathcal{C}} \left[w(c, \mathbf{p}) - \Gamma(c) + \varepsilon_c \right] \tag{6}$$

where

$$w(c, \mathbf{p}) = \max_{\mathbf{d} \in \mathcal{D}_c} \max_{\mathbf{q} \in R^K_{\geq 0}} \left[\left(\boldsymbol{\mu}_{\mathbf{d}} - \alpha \mathbf{p}_{\mathbf{d}} \right)' \mathbf{q} - 0.5 \mathbf{q}' \mathbf{\Lambda} \mathbf{q} \right]. \tag{7}$$

$w(c, \mathbf{p})$ is the household's indirect utility function (i.e., the maximum utility from a choice of $(\mathbf{d}, \mathbf{q})$) given shopping choice c and prices $\mathbf{p}$.

To solve the decision problem, we first solve the outer maximization in (7) to determine the store-category choices $\mathbf{d}$ conditional on a shopping choice c. To do so, we need to compare which store in c has a higher linear term in (5) for each category k :

$$\begin{aligned}\mathbf{d}(c, \mathbf{p}) &= [d_1(c, \mathbf{p}), \ldots, d_K(c, \mathbf{p})] \\ &= \left[\arg\max_{j \in c} \left(\mu_{j1} - \alpha p_{j1}\right), \ldots, \arg\max_{j \in c} \left(\mu_{jK} - \alpha p_{jK}\right)\right].\end{aligned} \tag{8}$$

The inner maximization problem in (7) is a quadratic programming in which, conditional on store-category choice $\mathbf{d}$ and prices $\mathbf{p}$, we need to find K category demands $\mathbf{q}$ under non-negative constraints:

$$\mathbf{q}(\mathbf{d}, \mathbf{p}) = \arg\max_{\mathbf{q} \in R^K_{\geq 0}} \left[\left(\boldsymbol{\mu}_{\mathbf{d}} - \alpha \mathbf{p}_{\mathbf{d}}\right)' \mathbf{q} - 0.5 \mathbf{q}' \boldsymbol{\Lambda} \mathbf{q}\right]. \tag{9}$$

We plug back the optimal $\mathbf{q}$ and $\mathbf{d}$ to (5). and rewrite the indirect utility of (6) as:

$$w(c, \mathbf{p}) = \left(\boldsymbol{\mu}_{\mathbf{d}(c,\mathbf{p})} - \alpha \mathbf{p}_{\mathbf{d}(c,\mathbf{p})}\right)' \mathbf{q}(c, \mathbf{p}) - 0.5 \mathbf{q}(c, \mathbf{p})' \boldsymbol{\Lambda} \mathbf{q}(c, \mathbf{p})$$

where $\mathbf{q}(c, \mathbf{p}) = \mathbf{q}(\mathbf{d}(c, \mathbf{p}), \mathbf{p})$.

Finally, after the indirect utility $w(c, \mathbf{p})$ is solved, we can derive the optimal shopping choice c. To do so, we solve the discrete choice problem in (6) given the distributional assumption on ε_c. The choice of c can be expressed as a list of choice probabilities for all possible shopping options in $\mathcal{C}$ given $w(c, \mathbf{p})$ and $n(c)$.

5.3 Heterogeneity

We explain in this section how tastes $(\boldsymbol{\mu}, \alpha, \boldsymbol{\Gamma})$ vary across households and weeks. We now introduce household and time subscripts, i and t respectively.

We begin with household i's taste at time t at store j for category k which is written in terms of observed and unobserved taste-shifters:

$$\mu_{itjk} = \xi_{fk} + \beta_{0k} \left(\beta_1 hz_i + \beta_2 sz_j + \mathbf{T}_t \boldsymbol{\beta}_T + \sigma_1 \nu^\mu_i + \sigma_2 \nu^\mu_{it} + \sigma_3 \nu^\mu_{ik} + \sigma_4 \nu^\mu_{ijk}\right) \tag{10}$$

Household taste μ_{itjk} consists of the chain type-category effect ξ_{fk}, common to all consumers, and reminder term $\beta_{0k} (\beta_1 hz_i + \ldots)$ that captures the deviation from the mean chain type-category effect.

The chain type-category effect ξ_{fk} can vary by chain type (f) and category (k). We use this term to capture the systematic difference in the provision of products and the product characteristics across different chain types under our definition. Particularly, traditional stores could provide homemade products that are not labeled.

The deviation term can vary across different household i, store j, and time t. It depends on observable characteristics, such as household size, store employment size, and time effects. The household size hz_i makes the utility greater for larger households. The employment size sz_j allows stores with more employees to offer greater utility (e.g., because of a better guarantee of food security). Quarter and year dummies T_t allow for seasonal and year effects. We also allow for unobserved heterogeneity in the utility. We include four random taste shocks (each i.i.d. $N(0,1)$): a household effect ν_i^μ , a household-time effect ν_{it}^μ , a household-category effect ν_{ik}^μ, and a household-store-category effect ν_{ijk}^μ. Their spreads are captured by $(\sigma_1, \sigma_2, \sigma_3, \sigma_4)$. These ν-terms introduce horizontal product differentiation at the store-category level, enabling each household to perceive stores differently given a store-category.

The price coefficient α_i is set to be heterogeneous across households according to their income y_i, household size hz_i, and unobserved characteristics ν_i^α:

$$\alpha_i = \left(\alpha_1 + \alpha_2 / \left(y_i / hz_i\right)\right) \nu_i^\alpha \tag{11}$$

The term ν_i^α is a Rayleigh(1) random shock which introduces heterogeneity in a parsimonious way while ensuring positive price sensitivity $\alpha_i > 0$ for all i, as long as α_1 and α_2 are positive.

Shopping costs are expressed as the sum of two components, one from $n(c)$, the number of stores in a trip specified by the shopping choice c, and total travel distance $dist_{ic} = 2\sum_{j \in c} dist_{ij}$ in the trip:

$$\Gamma_i(c) = \left(\gamma_{11} + \gamma_{12}\nu_{i1}^\Gamma\right) 1[n(c) = 2] + \left(\gamma_{21} + \gamma_{22}\nu_{i2}^\Gamma\right) dist_{ic}. \tag{12}$$

We also allow for unobserved household characteristics to influence the utility by adding ν_{i1}^Γ and ν_i^Γ, which are distributed i.i.d. $N(0,1)$.

It is important to note that Mexican households are spatially heterogeneous in terms of their choice sets and, therefore, the mix of products offered by the stores in their choice sets. Such heterogeneity can be both across households and within households due to the entry of stores owned by Walmart. These variations in the data are captured by the heterogeneity in the variety of stores and categories of groceries available to the households (as reflected by each i's choice set and $n_(c)$), characteristics of stores (e.g., employment size sz_j, and $dist_{ic}$) and the products (e.g., price $\mathbf{p}$) that are provided.

Specific to the application in our context, we can see from the model how spatial exposure in Walmart-owned stores and the products they provide can link with households' choices of grocery and store and how it is intertwined with the introduction of the sin-food tax. To see these linkages, consider the demand for grocery conditional on store-category choice $\mathbf{d}_{it}$ of household i in t. If $\mathbf{d}_{it}$ is such that the household chooses store $j \in \mathcal{J}_{it}$ for category k, then we can express grocery demand to category k, namely q_{itk}, using an implicit equation as [4]

$$q_{itk}(\mathbf{d}_{it}, \mathbf{p}_{it}) = \max\left[\frac{1}{\Lambda_{kk}}\left(\mu_{itjk} - \alpha_i p_{itjk} - 0.5\sum_{k' \neq k} \Lambda_{kk'} q_{itk'}(\mathbf{d}_{it}, \mathbf{p}_{it})\right), 0\right], \text{for the } j \in \mathcal{J}_{it} \text{ chosen for } k \tag{13}$$

[4]In equation 13, technically, the category of groceries that are available to the household i can also be influenced by which store j that i considers. For example, a store in i's choice set might not sell a certain category. For the parsimony, we do not reflect this dependence in the equation.

given the price $\mathbf{p}_{it}$ and store-category choice $\mathbf{d}_{it}$. The diagonal second-order quadratic terms, i.e., $\Lambda_{kk'}$ for any k, scale demand and (since α_i is fixed across categories) allow own-price effects to vary across categories. Suppose k is a grocery type representing low-calorie content in this equation. Suppose j is a Walmart-owned store that is available at time t but not in previous periods, which occurs if Walmart-owned stores enter the choice set, $\mathcal{J}_{it}$, of the household in time t. Then equation (13) illustrates that, as a Walmart-owned store enters into a household i's choice set in t, it influences household's demand for low-calorie grocery by affecting the store-category choice $\mathbf{d}_{it}$ (which includes Walmart after its entry), and repeatedly, all the store-specific heterogeneity in μ_{itjk} and p_{itjk} such that j is the Walmart-owned store available in time t. Given this equation, it is clear that the sin-food tax can influence households' grocery demand for such grocery products according to their exposure to Walmart-owned stores, as seen in the data. This is captured by the model in which households with or without Walmart-owned stores face differential changes in the price of sin-food in their neighborhood. In the interest of space, we refer the reader to Thomassen et al. (2017) for more analysis made available by the model.

6 Structural Estimation

In this section, we explain the details of implementing our structural estimation. We start by introducing the measurement for the key variables and primitives in the model. We then introduce whether the key model assumptions are satisfied by the data and how we select the households for the structural estimation. Finally, we present our estimation procedure.

6.1 Measurement

As is aligned with the model, each i is a household that can be uniquely identified in the data, and each t maps to a unique week. For constructing grocery categories, we subsume products into eight categories according to their definitive features and calorie contents. To make the model tractable, we constrain households' choice set to contain only the nearest 30 stores within 20 miles.[5] This makes $|J| = 30$ and the number of all the possible pairs of stores and singleton store, i.e., $|\mathcal{D}_c|$, to be $\frac{30\times 29}{2} = 465$. We elaborate below our practices in the measurement of each key model component.

Categories of Groceries k We group all the observed products in our data into five broad categories: Self-care and toiletries, household goods, beverages, dry foods, and dairy-based products[6]; to incorporate a healthiness dimension into our structural analysis, we further split each of the last three categories (which comprise all foods and beverages) into healthy and unhealthy. This leaves us with eight final categories: Self-

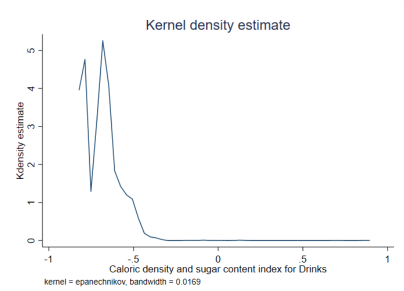
[5] We find a small proportion of households with less than 30 stores within 20 miles. We assign virtual stores to these households so their choice sets contain 30 stores. And the virtual stores are set to have unrealistically high prices in all the categories they provide.

[6] In general lines our categories follow Thomassen et al. (2017) however we do not observe purchases of meat, fruit and vegetables. Furthermore, we do not observe purchases of non-packaged bakery products; thus, we merge the bakery and dry grocery categories.

care and toiletries, household goods, healthy beverages, unhealthy beverages, healthy dry foods, unhealthy dry foods, healthy dairy-based products, and unhealthy dairy-based products.

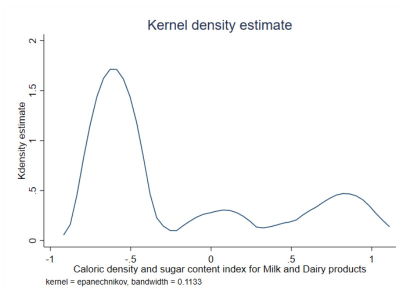
While we have data on the exact nutritional content of all products, the main outcome of interest in this study is caloric intake; moreover, the tax was defined based on caloric density (which implies that foods were not taxed according to their size) and sugar content. For these two reasons, we construct an index based on principal components analysis performed on the caloric density and sugar content of each product. With the index in hand, we define all products with an index above the 75th within-category percentile as unhealthy and the remainder as healthy.

Figure 8



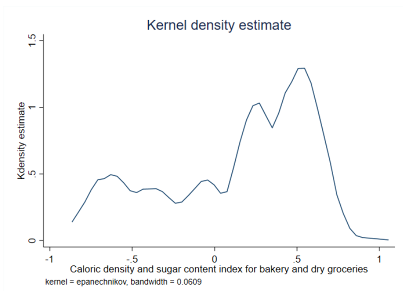
Kernel density estimate for the distribution of the calorie intensity and sugar content index among beverages.

Figure 9



Kernel density estimate for the distribution of the calorie intensity and sugar content index among milk and dairy products.

Figure 10



Kernel density estimate for the distribution of the calorie intensity and sugar content index among bakery and dry grocery.

Price Index p Having defined product categories, we are interested in how prices and availability at the store level affect how households adapt to relative prices of healthy foods and beverages. Therefore, we construct a price index at the firm level. Specifically, we make our price variable following the two-level procedure of Thomassen et al. (2017). In the first step, the goal is to capture intra-store preferences for specific types of products within the broader categories; this ensures that the aggregate price index accurately reflects the price of goods more frequently purchased within the store. In the second step, the aggregation is done based on total revenue per product group (without considering store level variation) so that product type-store combinations not frequently purchased are not over-represented in the final index. For specific details on the imputation algorithms, see the appendix B.

Chain Types f As is emphasized in introducing the chain type-category effects, i.e., ξ_{fk}, we need to properly group chains into chain types. The purpose of the grouping is to let the term ξ_{fk} account for the common features shared by all chains of the same type and category. Therefore, chains in the same group should reveal common chain and store characteristics. Therefore, we use a cluster analysis to group chains into chain types. The variables we relied on were the mean and standard deviation of the price indices per category, the median market share and count of products (measured at the city level) for each category, the median store size (as measured by number of employees reported in DENUE) and the total number of cities in which the store is observed. We argue that this variable summarizes the size and comparative advantages regarding firms' prices and product variety. Finally, we group all chains into eight chain types. They include traditional corner stores, other traditional stores, small chains, medium chains, and big chains except for Walmart-owned stores, Walmart Super Center, Aurrera (smaller Walmart stores targeted at low-income households), and Aurrera Express (also owned by Walmart).

Assignment of Stores and Imputation of Certain Stores' Characteristics To analyze households' purchasing patterns, we need to know which stores are within a feasible distance and, therefore, likely to be considered; combining data on the exact location of chain stores with the location of the neighborhoods in which households live, we construct household-specific choice sets based on distance. For this procedure, we matched all the stores that appear in DENUE with all the chains in which at least one purchase was observed in the Kantar data. With this procedure, we can find locations for approximately 76% of all the chains that appear in the Kantar data, which account for about 90% of observed expenditures. Following this procedure, we can match 80% of the total expenditures of the average household in our data to stores that lie within less than 5 kilometers of the household's neighborhood. Having distance from every household to every store, we define a household's choice set as the set of all stores that lie within a 20-kilometer distance.[7] More details and results of this procedure are provided in appendix C.

Household Size hz **and Employment Size** sz The information on the demographics of households can be found in the home scanner dataset. However, the dataset does not provide household income. Instead, they provide the socioeconomic group each household belongs to, ranging from 1 to 6. With this information, we impute household income levels for each socioeconomic group according to a public report that maps the average household income during the sample period for each socioeconomic group. As for the employment size information, it is available in the DENUE dataset. However, it only allows the number of employees to fall into six exclusive intervals of employee size (e.g., "having less than five employees", "having 10 to 100 employees). We use the logarithm of the lower bound of such intervals as the employment size measure for stores.

Quantities of Groceries q To get the quantities of grocery demand at the category level as the model requires, we first aggregate expenditure to the store-category-week level for each household and divide by price index as constructed.

6.2 Sample Selection and Supporting Evidence for Key Model Assumptions

We jointly inspect the supporting evidence for key model assumptions for traceability and select specific households from the full sample for the structural estimation. We eliminate households from the full sample and do not consider them as the candidates to enter the sample for structural estimation. We then construct a subsample from the remaining households by randomly selecting households and their weekly purchase data. We elaborate on the implementation details below.

[7]Note that since Kantar does not record expenditures at the store level but only at the chain level when two or more stores that belong to the same chain belong to the choice set of a household, we only keep the closest one.

In the first step of elimination, we drop households that do not have long enough panels of weekly purchase data. Primarily, the information provided by these households may be noisy and low-quality.[8] Further, including households with too short panels stops us from using the timing of purchases, which brings temporal variation in prices and store availability. Practically, we need the households selected for the structural estimation to have sufficiently distant weeks to facilitate further randomization of households.

In the second step of elimination, besides these households with short panels, we also eliminate some households based on their behaviors. In this step, we concurrently inspect whether households' purchase patterns largely satisfy our assumptions for tractability. As shown in appendix A, although most households meet the behavioral patterns that support the key model assumptions, some deviate from the patterns. Therefore, we drop them from the eligible households that could enter the structural estimation. We document in the appendix A the data loss in each elimination step and the summary statistics of the remaining samples surviving each elimination. Since the remaining households have similar summary statistics with the full sample, we believe these "eligible" households are reasonable representatives of the full sample. Finally, all the elimination steps leave us with around 7920 households. We randomly pick households further for the structural estimation because the model is computationally costly if we use all the eligible households. We select a subsample, a panel of 2,000 consumers, and three weeks per household, following the randomization procedure applied in Thomassen et al. (2017).

6.3 Estimation Strategy

To be written.

7 Results

In-progress

References

[1] Arturo Aguilar, Emilio Gutierrez, and Enrique Seira. "The effectiveness of sin food taxes: evidence from Mexico." In: *Instituto Tecnológico Autónomo de México (ITAM) Working Paper* (2019).

[2] Hunt Allcott et al. "Food deserts and the causes of nutritional inequality." In: *The Quarterly Journal of Economics* 134.4 (2019), pp. 1793–1844.

[3] David Atkin, Benjamin Faber, and Marco Gonzalez-Navarro. "Retail globalization and household welfare: Evidence from mexico." In: *Journal of Political Economy* 126.1 (2018), pp. 1–73.

[4] Emek Basker. "Job creation or destruction? Labor market effects of Wal-Mart expansion." In: *Review of Economics and Statistics* 87.1 (2005), pp. 174–183.

[8] For example, these households are surveyed by KANTAR, yet although they do not fully understand how to use the expenditure recorder device properly, their recorded data may not be trustworthy.

[5] May A Beydoun, Lisa M Powell, and Youfa Wang. “The association of fast food, fruit and vegetable prices with dietary intakes among US adults: is there modification by family income?” In: *Social science & medicine* 66.11 (2008), pp. 2218–2229.

[6] John Cawley. “An economy of scales: A selective review of obesity’s economic causes, consequences, and solutions.” In: *Journal of health economics* 43 (2015), pp. 244–268.

[7] Chanjin Chung and Samuel L Myers Jr. “Do the poor pay more for food? An analysis of grocery store availability and food price disparities.” In: *Journal of consumer affairs* 33.2 (1999), pp. 276–296.

[8] M Arantxa Colchero et al. “Beverage purchases from stores in Mexico under the excise tax on sugar sweetened beverages: observational study.” In: *bmj* 352 (2016), h6704.

[9] Joan Costa-Font and Núria Mas. “‘Globesity’? The effects of globalization on obesity and caloric intake.” In: *Food Policy* 64 (2016), pp. 121–132.

[10] Ronald Cotterill and Andrew Franklin. *The urban grocery store gap.* Tech. rep. 1995.

[11] Charles Courtemanche and Art Carden. “Supersizing supercenters? The impact of Walmart Supercenters on body mass index and obesity.” In: *Journal of Urban Economics* 69.2 (2011), pp. 165–181.

[12] Janet Currie et al. “The effect of fast food restaurants on obesity and weight gain.” In: *American Economic Journal: Economic Policy* 2.3 (2010), pp. 32–63.

[13] David M Cutler, Edward L Glaeser, and Jesse M Shapiro. “Why have Americans become more obese?” In: *Journal of Economic perspectives* 17.3 (2003), pp. 93–118.

[14] Pierre Dubois, Rachel Griffith, and Aviv Nevo. “Do prices and attributes explain international differences in food purchases?” In: *American Economic Review* 104.3 (2014), pp. 832–67.

[15] Jason M Fletcher, David E Frisvold, and Nathan Tefft. “The effects of soft drink taxes on child and adolescent consumption and weight outcomes.” In: *Journal of Public Economics* 94.11-12 (2010), pp. 967–974.

[16] Osea Giuntella, Matthias Rieger, and Lorenzo Rotunno. “Trade in unhealthy foods and obesity: Evidence from Mexico.” In: (2017).

[17] Tadeja Gracner. “Bittersweet: how prices of sugar-rich foods contribute to the diet-related disease epidemic in Mexico.” In: *Job Market Paper* (2015).

[18] Michael Grossman, Erdal Tekin, and Roy Wada. “Food prices and body fatness among youths.” In: *Economics & Human Biology* 12 (2014), pp. 4–19.

[19] Matthew Harding and Michael Lovenheim. “The effect of prices on nutrition: comparing the impact of product-and nutrient-specific taxes.” In: *Journal of health economics* 53 (2017), pp. 53–71.

[20] Jerry Hausman and Ephraim Leibtag. “Consumer benefits from increased competition in shopping outlets: Measuring the effect of Wal-Mart.” In: *Journal of Applied Econometrics* 22.7 (2007), pp. 1157–1177.

[21] James O Hill, Holly R Wyatt, and John C Peters. “Energy balance and obesity.” In: *Circulation* 126.1 (2012), pp. 126–132.

[22] Thomas J Holmes. “The diffusion of Wal-Mart and economies of density.” In: *Econometrica* 79.1 (2011), pp. 253–302.

[23] Stefan Hut and Emily Oster. “Changes in Household Diet: Determinants and Predictability.” In: (2015).

[24] Leonardo Iacovone et al. “Supplier responses to Walmart’s invasion in Mexico.” In: *Journal of International Economics* 95.1 (2015), pp. 1–15.

[25] Beata S Javorcik and Yue Li. “Do the biggest aisles serve a brighter future? Global retail chains and their implications for Romania.” In: *Journal of International Economics* 90.2 (2013), pp. 348–363.

[26] Beata Javorcik, Wolfgang Keller, and James Tybout. *Openness and industrial response in a Wal-Mart world: a case study of Mexican soaps, detergents, and surfactant producers.* The World Bank, 2006.

[27] Panle Jia. “What happens when Wal-Mart comes to town: An empirical analysis of the discount retailing industry.” In: *Econometrica* 76.6 (2008), pp. 1263–1316.

[28] Phil R Kaufman. “Rural poor have less access to supermarkets, large grocery stores.” In: *Rural America/Rural Development Perspectives* 13.2221-2019-2662 (1998), pp. 19–26.

[29] Phillip R Kaufman et al. *Do the poor pay more for food? Item selection and price differences affect low-income household food costs.* Tech. rep. 1997.

[30] Dragan Miljkovic et al. “Globalisation and obesity.” In: *The World Economy* 38.8 (2015), pp. 1278–1294.

[31] Kimberly Morland et al. “Neighborhood characteristics associated with the location of food stores and food service places.” In: *American journal of preventive medicine* 22.1 (2002), pp. 23–29.

[32] Lisa Oberlander, Anne-Célia Disdier, and Fabrice Etilé. “Globalisation and national trends in nutrition and health: A grouped fixed-effects approach to intercountry heterogeneity.” In: *Health economics* 26.9 (2017), pp. 1146–1161.

[33] World Health Organization. *The world health report 2002: reducing risks, promoting healthy life.* World Health Organization, 2002.

[34] Lisa M Powell and Frank J Chaloupka. “Food prices and obesity: evidence and policy implications for taxes and subsidies.” In: *The Milbank Quarterly* 87.1 (2009), pp. 229–257.

[35] Donald Rose and Rickelle Richards. “Food store access and household fruit and vegetable use among participants in the US Food Stamp Program.” In: *Public health nutrition* 7.8 (2004), pp. 1081–1088.

[36] Øyvind Thomassen et al. “Multi-category competition and market power: a model of supermarket pricing.” In: *American Economic Review* 107.8 (2017), pp. 2308–51.

[37] Roberto De Vogli et al. “Economic globalization, inequality and body mass index: a cross-national analysis of 127 countries.” In: *Critical Public Health* 24.1 (2014), pp. 7–21.

[38] Richard Volpe, Abigail Okrent, and Ephraim Leibtag. “The effect of supercenter-format stores on the healthfulness of consumers’ grocery purchases.” In: *American Journal of Agricultural Economics* 95.3 (2013), pp. 568–589.

[39] Zy Weinberg. *No Place to Shop: The Lack of Supermarkets in Low-income Neighborhoods: Analysis of a University of Connecticut Study on Low-income Communitites and Households Receiving Public Assistance in 21 Metropolitan Areas*. Public Voice for Food and Health Policy, 1995.

A Sample Selection for Structural Estimation

We conduct three steps of elimination on the full sample. The first elimination aims to eliminate households with lower-quality observations, and the second and third eliminations instead target the households that severely violate the model assumptions relative to other households. Denote household expenditure as E. We summarize household expenditure to the household-week-store-category level, i.e., E_{itjk}. For ease of exposition, we use E_i, E_{it}, E_{itj}, E_{itk} to represent the expenditure summed at the corresponding levels (e.g., $E_{it} = \sum_j \sum_k E_{itjk}$).

There are three assumptions made for tractability. The second and third steps of elimination center on making the remaining households meet these assumptions relatively more. These assumptions are:

- Assumption 1: Households' choice sets are the nearest 30 stores to their residential locations.
- Assumption 2: For each i in each week t, at most 2 stores are visited.
- Assumption 3: For each i in each week t, if the purchase behaviors take place in two stores, then for each k purchased, only one store is visited

The KANTAR dataset includes 14116 unique households ranging from 2012 to 2015.

A.1 First elimination

In the first elimination, we eliminate households with short weekly expenditure records or poor data quality. Truncating these households helps us in two ways. First, we are concerned that households with short purchase records bring too many measurement errors. For example, these households may not fully understand how to properly use the expenditure recorder device. Second, as in Thomassen et al. (2017), we need to randomly select households' purchases at different weeks for the estimation sample, and therefore, the available weeks we choose from need to be sufficiently distant from each other within a household. This is particularly important to us since we expect to involve the analysis of Walmart's entry on purchase.

In practice, we drop households if they survive in the KANTAR weekly expenditure dataset for less than two quarters or have less than 24 weeks of purchase records. In this step, 21% (2950/14116) of the households are dropped. Thomassen et al. (2017) also drops households with too short records to conduct further analysis, resulting in 23% of the households dropping from their sample.[9]

There are 1596 households from 2012 to 2015 whose identifiers appear in the purchase records while not in the dataset of household characteristics. Most such households were newly entered into the KANTAR expenditure dataset in 2015. Therefore, we drop the union of these households and the above households with a shorter sample length.

[9] This is the only considerable elimination Thomassen et al. (2017) does because the three assumptions are largely satisfied in their case.

In the first elimination, we are left with 10378 households. These households have relatively long purchase records and no problems with missing variables. In the following elimination steps, all the descriptive statistics we provide are conditional on the survivors of the first elimination.

A.2 Second elimination

In the second elimination, we delete households whose total expenditures are large outside the 30 nearest stores of their choice set. We show below the percentile of the total expenditure spent within the 30 nearest stores across all households. The table shows that more than 90% of households spend more than 83.9 % within their 30 nearest stores. Therefore, assumption one is plausible given the majority of households spend most of their expenditures within the 30 nearest stores.

Therefore, in practice, we drop a household i if $\frac{\sum_{j:30nearest} E_{ij}}{E_i} < 0.9$. This results in 1562 households (15%) dropping from 10378 households, leaving us 8816 households.

A.3 Third elimination

In the third elimination, we inspect assumptions 2 and 3. In KANTAR, among expenditures of all pairs of household-week $i-t$, roughly 46% of them are spent on less than three stores and 47% of them on three or four stores. Hence, the frequency of the appearance of weekly expenditures spent in more than two stores is high, which makes assumption 2 seem implausible. However, assumption 2 can still be reasonable given the following descriptive statistics. We show the expenditure share from the two stores where households spend the most relative to the overall expenditure of that weekly purchase for all households. In other words, it is defined as the expenditure share of the highest and the second highest E_{itj} across all js of the i, t. The table below shows the distribution of such statistics over more than 1.4 million pairs of $i-t$. It shows that more than 90% of the $i-t$ pairs involve two major stores, in which they account for more than 73.3% of the total $i-t$ expenditure.

Continuing to inspect the plausibility of assumption 3, we provide the distribution of the higher expenditure share spent among the two stores for each $i-t$ and category k. For example, suppose we see that in a weekly expenditure profile indexed by i and t, the expenditure E on category k is distributed as 64% on one store j_1 and 36% on the other store j_2. In that case, we record the number 0.64 for the trio $i-t-k$. We show the percentile of the distribution of these statistics below across all trios of $i-t-k$. The 10th percentile is 0.63, indicating that 90% of the expenditures E_{itk} involve one store taking more than 63% of the expenditure on the category k. Moreover, the 50th percentile is 1, meaning 50% of the E_{itk} involves expenditure on one store.

Therefore, in practice, to implement the third elimination, we first tag each E_{it} as "healthy" if it satisfies

$$\frac{E_{itj_1} + E_{itj_2}}{E_{it}} > 80\%,$$

stats	share~30
N	10378
mean	.9458438
sd	.1201296
p5	.7072771
p10	.8392031
p25	.9558496
p50	.9933703
p75	.9995903
p95	1

Figure A1: Distribution of share-within-30-nearest-stores among 10378 households. In the table, "px" means x-percentile of the statistics.

stats	max2ex~e
N	1467309
mean	.9081419
sd	.1147586
p5	.6740947
p10	.7338018
p25	.8396226
p50	.9563949
p75	1
p95	1

Figure A2: The distribution of the share of the two mostly-spent stores among all 1.4 million pairs is i in t. In the table, "px" means x-percentile of the statistics.

where j_1 and j_2 indicates the first and the second highest stores in which i spends money on in t. Similarly, we tag each E_{itk} as "healthy" if

$$\frac{E_{it\tilde{j}k}}{E_{itk}} > 90\%,$$

where $\tilde{j}$ means the store in which the expenditure k household i in t spends on is the highest. Therefore, a "healthy" expenditure record can be understood as being closer to satisfying the assumptions 2 or 3, respectively, compared to an unhealthy expenditure record. Further, each household gets a measure of the degree of "healthiness" according to the two criteria. We then drop the union of the households with the lowest 5% healthiness in both measures, resulting in around 500 households loss.

Therefore, in the end, all the elimination leaves us with around 7920 households. These households are the candidates for constructing the structural estimation sample. We continue to trim all candidates' weekly expenditure records to satisfy the three assumptions exactly. To do this, firstly, all expenditures outside of the 30 nearest stores are coded as zero. Secondly, their expenditures of all purchases outside the two stores where they spend the most are coded as zero. Thirdly, in each weekly expenditure record, the expenditures spent on each store category with a lower share are coded as zero. From the elimination steps, the trimming does not severely distort the information stored in these households' expenditures.

A.4 Sample comparison

We compare samples remaining under each elimination with the full sample in the mean and variance of household characteristics. Across samples, they are balanced in the selected list of household characteristics.

stats	chainr~e
N	8029855
mean	.9222415
sd	.1505693
p5	.5588008
p10	.6423144
p25	.975265
p50	1
p75	1
p95	1

Figure A3: The distribution of the higher expenditure share of the two stores within a category among around 8 million pairs purchase-category.

	(1) Full sample	(2) First survv.	(3) Secon..	(4) Third Survv.
age_hw	41.21	41.67	41.49	41.34
	(14.19)	(14.03)	(14.03)	(14.10)
emp_hw	0.26	0.25	0.25	0.25
	(0.437)	(0.432)	(0.433)	(0.433)
edu_hw	5.18	5.13	5.04	5.05
	(2.674)	(2.647)	(2.612)	(2.617)
age_head	43.77	44.11	43.94	43.76
	(14.04)	(13.91)	(13.90)	(13.99)
emp_head	4.01	4.03	4.01	4.01
	(1.328)	(1.336)	(1.313)	(1.320)
edu_head	5.59	5.55	5.46	5.45
	(3.036)	(3.074)	(3.064)	(3.083)
nse_loc	3.74	3.74	3.78	3.79
	(1.373)	(1.381)	(1.368)	(1.370)
n_member	4.22	4.28	4.27	4.24
	(1.764)	(1.773)	(1.779)	(1.767)
n_adult	2.54	2.57	2.56	2.54
	(1.107)	(1.106)	(1.104)	(1.092)
n_adol	1.55	1.59	1.59	1.58
	(1.120)	(1.122)	(1.125)	(1.125)

Figure A4: Comparison of household characteristics of the samples (to be continued).

n_female	2.25	2.29	2.28	2.27
	(1.195)	(1.203)	(1.202)	(1.202)
n_male	1.95	1.98	1.98	1.96
	(1.216)	(1.218)	(1.225)	(1.218)
prop_car	0.50	0.51	0.50	0.49
	(0.644)	(0.647)	(0.642)	(0.643)
prop_compu~r	0.44	0.46	0.44	0.44
	(0.707)	(0.729)	(0.720)	(0.714)
prop_tv	1.58	1.58	1.57	1.56
	(0.973)	(0.981)	(0.963)	(0.952)
health_hea~i	27.56	27.71	27.69	27.62
	(5.417)	(5.268)	(5.272)	(5.204)
health_hea~c	0.04	0.04	0.04	0.04
	(0.199)	(0.194)	(0.192)	(0.189)
health_hea~m	27.36	27.46	27.42	27.35
	(4.961)	(4.848)	(4.832)	(4.724)
N	15733	10365	8788	7920

Figure A5: (Continued)Comparison of household characteristics of the samples. Variables are age, employment status, education level of housewife and head, socioeconomic status at the location (nse_loc), the number of family members of different kinds, the property holdings (car, computer, TV), and health status(two body mass indices, diabetes rate)

B Price index construction

Purchases are observed in the raw data at the *store-household-date-bar code* level. To estimate the model, we need a price for every category (*Dry grocery, self-care products, healthy beverages, unhealthy beverages, healthy dry groceries, unhealthy dry groceries, healthy dairy, and unhealthy dairy*) at every store on every date and city. To aggregate from barcode level prices to category prices, we use an intermediate level, which we call *product group*. The definition of product groups is inherited from the KANTAR dataset, which defines comparable products. A barcode belongs to one and only one *product group* and a category (*ej:* 200 ml Head and Shoulders $\in$ Shampoo $\in$ Self Care)

The price variable is constructed in two steps: We first aggregate individual bar code prices into *product group-store* prices. And then, we aggregate *product group-store* prices into *category-store prices*.

The described procedure requires a price for all relevant *bar code-store* combinations every week. However, not all bar codes are purchased every week in all stores. Thus, we impute unobserved prices using the following procedure:

1. Using the purchase data, we created dummies for all possible bar code#Store#(City/Region/Mexico)#(Week/Quarter combinations. These combinations are possible imputation levels.

2. We regress observed prices on the obtained dummies. This fully saturated model can be seen as a non-parametric estimator of the price.

3. We sort imputation levels by predictive power (R^2).

4. We calculate the average and median bar code price at each imputation level.

5. We impute the median price to each bar code#Store#City#Week combination observed in the data from the available level with the highest explanatory power.

B.1 From bar codes to product groups

The first step is obtaining a price for each product category in every store and weekly. Let p_{swybgc} be the price in chain s of bar code b that belongs to product group g and category c during week w of year y. In this step, the prices are obtained from the imputation algorithm and are aggregated using weights that aim to capture intra-category and intra-store preferences.

If all bar codes were sold in every chain, the weight ω_{bg} assigned to bar code b within the product group g would be defined by its share of total observed sales (measured by volume) among barcodes that comprise g. However, not all chains sell all barcodes. Thus, the weights are computed only among chains that sell barcode b. This means that if $V_{gs(b)}$ is the total volume of product group g observed sales in chains $s(b)$ that sold bar code b at least once and V_b is the total volume of bar code b sold. ω_{bg} is computed as $\frac{V_b}{V_{bs(g)}}$.

The previous procedure leads to weights that do not add up to one within chain-product group combinations. Thus, weights are normalized according to the following rule:

$$\hat{\omega_{bg}} = \frac{\omega_{bg}}{\sum\limits_{b \in B_{gs}} \omega_{bg}}$$

Where B_{gs} is the set of barcodes in product group g sold by chain S.

The price for product group g of category c at chain s in week w is computed as:

$$p_{wysgc} = \sum_{b \in g} \hat{\omega_{bg}} p_{swybgc}$$

Finally, p_{wysgc} is normalized by the value of the price index during the first week we observe in our sample.

B.2 From product groups to categories

In the second step, we calculate weights according to total observed expenditures per product group. (Note that at this level, within-store preferences are not considered.) Let $z_g c$ be the total observed expenditures in product group g of category c and let z_c be the total observed expenditures in category c. Then, the price for category c in chain s during week w of year y is obtained as:

$$P_{cswy} = \sum_{g \in c} \frac{z_g c}{z_c} P_{wysgc}$$

In our data, we do not have a price for every product group and every store. In this case, we assume that the price was equal to the highest observed price for that product.

C Matching stores to households

We used string search to find the location of stores in DENUE for all the chains that appear in the Kantar dataset.

The Kantar dataset only records the chain when purchases occur at relatively large chains. This implies that a large proportion of households' expenditures can only be associated to a type of store but not to a chain.

Our first step is to link chains that appear in the Kantar dataset to stores that are registered in DENUE. This first step is done using string searches.

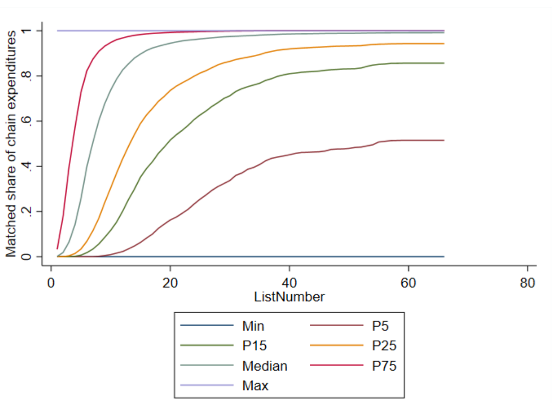
Then, for every household in the Kantar dataset we created a list comprising all the stores within a 20 kilometers radius around the household's location.

Then, for every household, we calculated the proportion of total observed chain expenditures that was captured by its associated list of stores. We found that 25 percent of households had a share of matched chain expenditures below .3. This implied that for 25 percent of households we were matching less than 65 percent of their total observed expenditures to either traditional (unmatcheable) retailers or chains successfully linked to at least one store in DENUE.

The main cause for this "pathological" households was that most of them appear only a few times in the purchases data (In less than three different weeks). Ignoring these households leads to a clear improvement in the quality of the match.

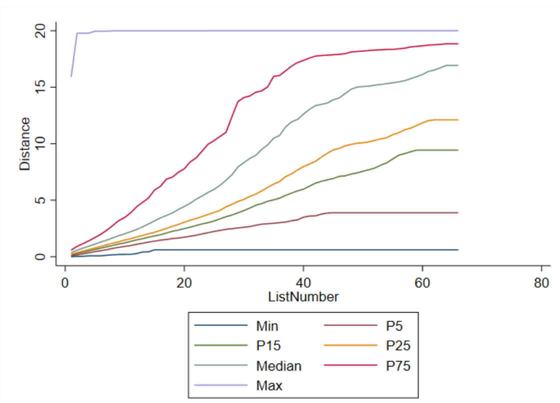
Quality of match by distance to households

Figure A6



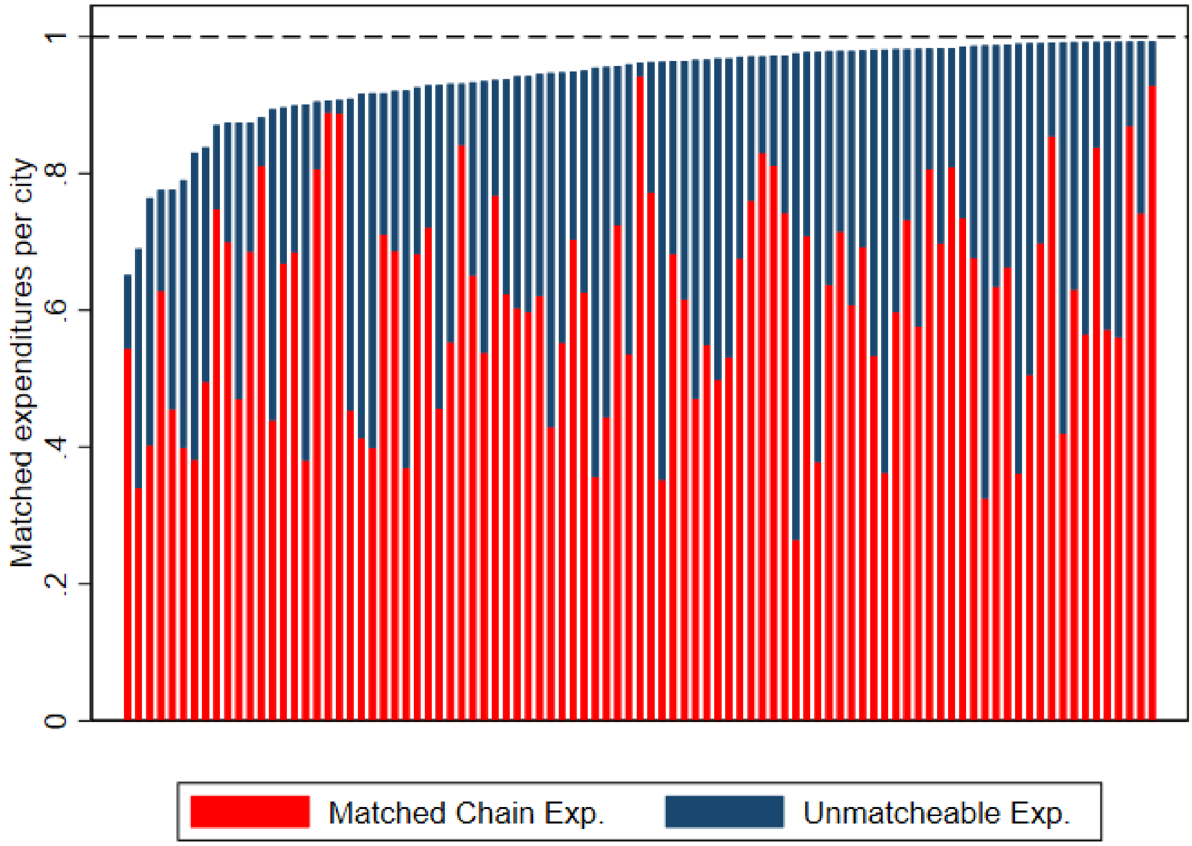
The x-axis represents stores sorted increasingly by distance relative to each household. The y-axis represents the accumulated share of matched chain store expenditures. The sample is restricted to households whose purchases were observed for at least three consecutive weeks.

Figure A7



The x-axis represents stores sorted increasingly by distance relative to each household. The y-axis represents the actual distance between stores and households. The sample is restricted to households whose purchases were observed for at least three consecutive weeks.

Figure A8



The figure reports matched chain expenditures within 20 kilometers and traditional expenditures as a proportion of total observed expenditures at the city level. Each bar corresponds to a city.