Online Appendix

Food Deserts and the Causes of Nutritional Inequality

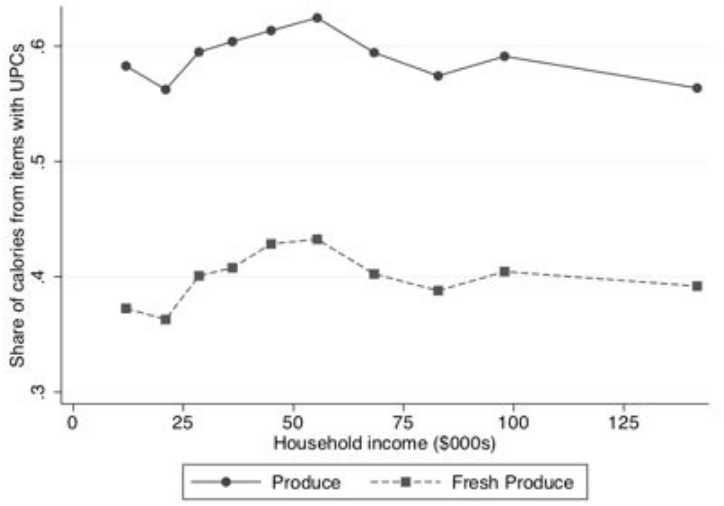
Hunt Allcott, Rebecca Diamond, Jean-Pierre Dubé, Jessie Handbury,

Ilya Rahkovsky, and Molly Schnell

A Appendix to Data Section

A.A Magnet Calorie Shares

Figure A1: **Magnet Data: Share of Produce from Packaged Items**



Notes: This figure uses the Nielsen Homescan "magnet" subsample for 2004–2006 to show the share of produce and fresh produce calories coming from packaged items with UPCs, which are the items that we observe outside the magnet subsample. The x-axis presents bins of average household income across all years the household is observed in Homescan. "Produce" includes fresh, dried, canned, and frozen produce. Observations are weighted for national representativeness.

A.B Health Index

Table A1: **Correlations Between Health Index and Its Components in Homescan**

Component	Correlation with Health Index
Adequacy ("healthy") components	
Total fruits	0.56
Whole fruits	0.57
Vegetables	0.41
Greens and beans	0.47
Whole grains	0.50
Dairy	0.25
Total protein	0.42
Sea and plant protein	0.64
Monounsaturated fat	0.11
Polyunsaturated fat	0.07
Moderation ("unhealthy") components	
Refined grains	-0.33
Sodium	-0.09
Added sugar	-0.41
Saturated fat	-0.21
Solid fats	-0.44

Notes: Using Homescan household-by-year data for 2004–2016, this table presents the correlation coefficients between the Health Index and its components, using components in units per 1,000 calories consumed. Observations are weighted for national representativeness.

Table A2: **Associations Between Health Outcomes and Dietary Quality Measures**

(a) **Dependent Variable: Body Mass Index**

	(1)	(2)	(3)	(4)
Healthy Eating Index (standard)	-1.272***	-1.251***		
	(0.066)	(0.400)		
Health Index (linearized)			-1.230***	-1.149***
			(0.065)	(0.365)
Controls	No	Yes	No	Yes
R^2	0.029	0.882	0.032	0.882
N	10,993	10,993	10,993	10,993
Dependent var. mean	29.47	29.47	29.47	29.47

(b) **Dependent Variable: Diabetic**

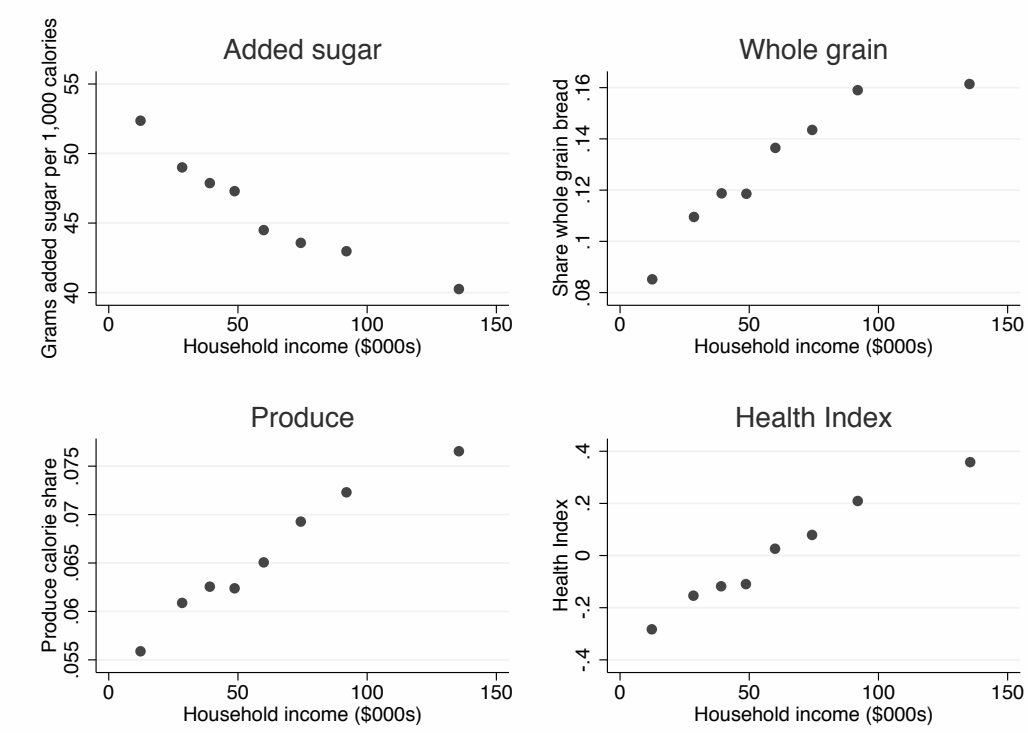
	(1)	(2)	(3)	(4)
Healthy Eating Index (standard)	-0.0265***	-0.0298		
	(0.0032)	(0.0196)		
Health Index (linearized)			-0.0297***	-0.0323*
			(0.0029)	(0.0173)
Controls	No	Yes	No	Yes
R^2	0.005	0.891	0.007	0.891
N	11,067	11,067	11,067	11,067
Dependent var. mean	0.17	0.17	0.17	0.17

Notes: This table presents regressions of health outcomes on dietary quality measures, using household-level Nielsen Homescan data. Body Mass Index is weight (in kilograms) divided by the square of height (in meters). Diabetic takes value 1 if the panelist reported that she had been diagnosed with diabetes, and 0 otherwise. If two household members responded to the PanelViews survey, we take the mean of each survey variable across the two respondents. Dietary quality measures are normalized to have a mean of zero and a standard deviation of one across households; we then take the calorie-weighted average across all years the household was observed in the Homescan sample. "Controls" are household-by-census tract fixed effects and household demographics (natural log of income, natural log of years of education, age indicators, household size, race indicators, a married indicator, employment status, and weekly work hours). Observations are not weighted for national representativeness. Robust standard errors, clustered by household, are in parentheses. *, **, ***: statistically significant with 10, 5, and 1 percent confidence, respectively.

B Appendix to Stylized Facts Section

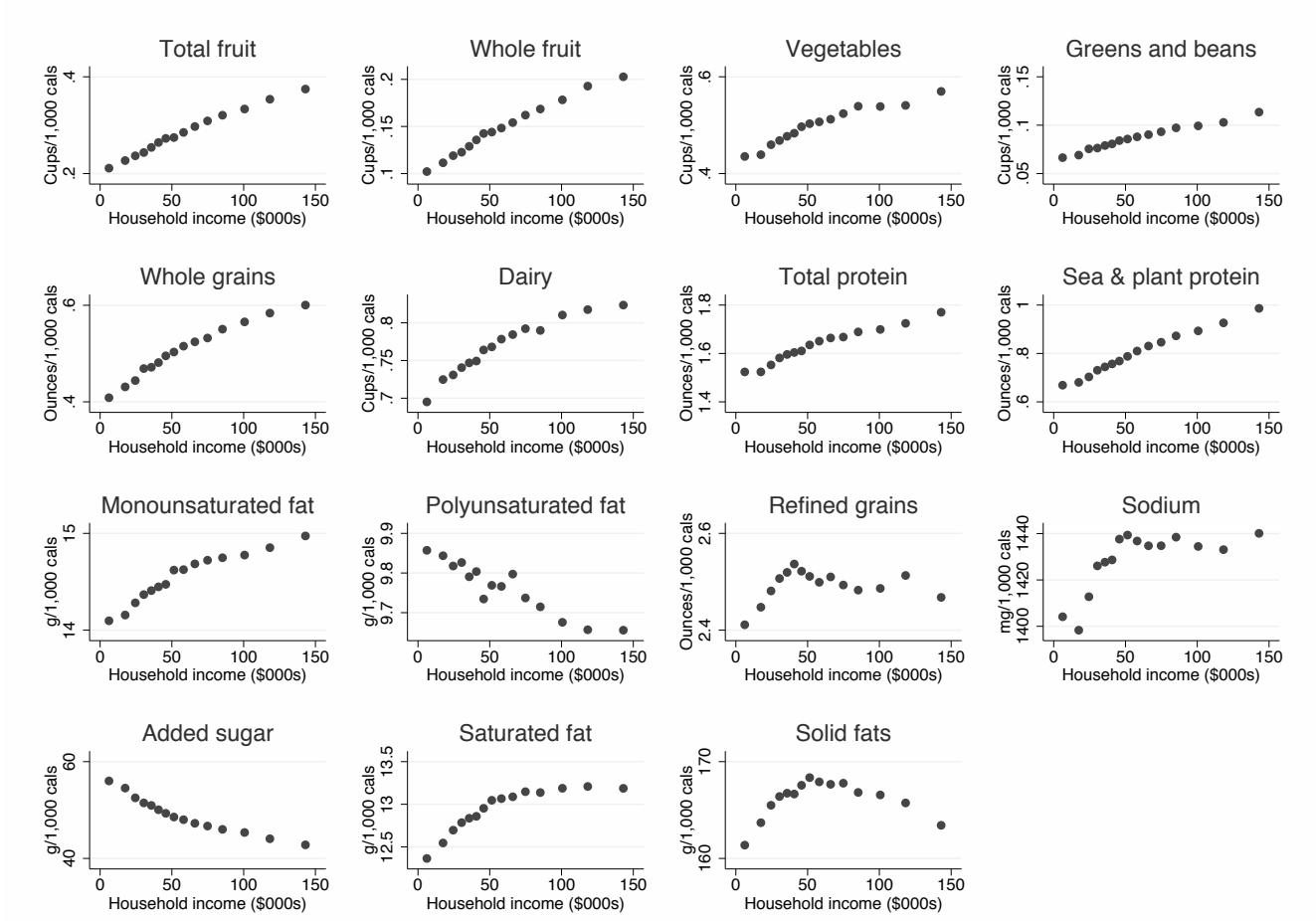
B.A Additional Figures and Tables

Figure A2: **Magnet Subsample: Healthful Purchases by Household Income**



Notes: This parallels Figure 1, except it uses the 2004–2006 magnet subsample, which also records purchases of non-UPC items such as bulk produce. Each panel presents a binned scatterplot of a grocery healthfulness measure against average household income across all years the household is observed in Homescan, residual of age and year indicators and household size. Added sugar is the grams of added sugar per 1,000 calories purchased; whole grain is the calorie-weighted average share of bread, buns, and rolls purchases that are whole grain; produce is the share of calories from fresh, canned, dried, and frozen fruits and vegetables; and the Health Index is our overall measure of the healthfulness of grocery purchases, normalized to have a mean of zero and a standard deviation of one across households. Observations are weighted for national representativeness.

Figure A3: **Healthy Eating Index Components by Household Income**



Notes: This figure presents Nielsen Homescan data for 2004–2016. Each panel presents a binned scatterplot of a dietary quality measure against average household income across all years the household is observed in Homescan, residual of age and year indicators and household size. The first 10 panels present the "healthy" dietary components of the Healthy Eating Index, while the final five panels present the "unhealthy" components. Observations are weighted for national representativeness.

Table A3: **Correlates of the Count of Produce UPCs Available in RMS Stores**

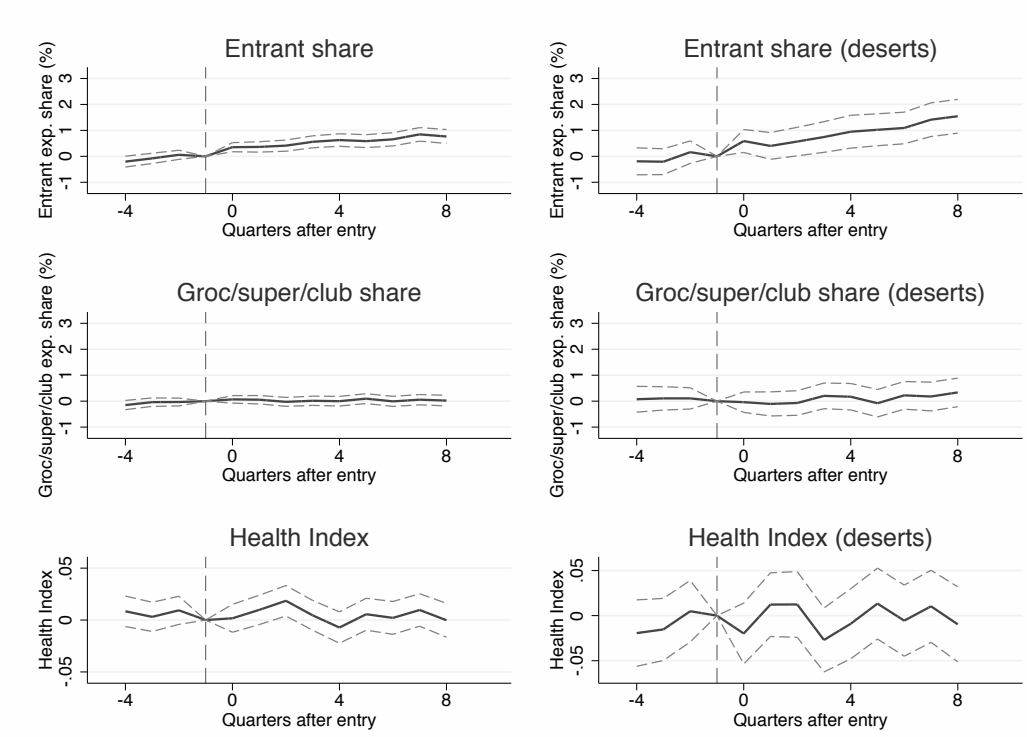
	(1)	(2)	(3)
ln(Zip median income)	0.405***	0.000	0.094***
	(0.009)	(0.005)	(0.003)
ln(Annual revenue)		0.366***	
		(0.001)	
1(Large grocery)			1.488***
			(0.004)
1(Small grocery)			1.036***
			(0.008)
1(Supercenter/club)			0.883***
			(0.012)
1(Convenience store)			-0.143***
			(0.002)
1(Drug store)			-0.083***
			(0.002)
Observations	369,903	369,903	369,903
R^2	0.04	0.80	0.92

Notes: This table presents regressions of the count of produce UPCs available in RMS stores on store characteristics, using 2006–2016 Nielsen RMS data at the store-by-year level. ln(Annual revenue) is revenue from packaged grocery items with UPCs. "Large" ("small") grocery stores are those with at least (less than) $5 million in annual revenue. Mass merchants other than supercenters and club stores are the omitted store type in column 3. Robust standard errors, clustered by zip code, are in parentheses. *, **, ***: statistically significant with 10, 5, and 1 percent confidence, respectively.

C Appendix to Reduced-Form Event Studies

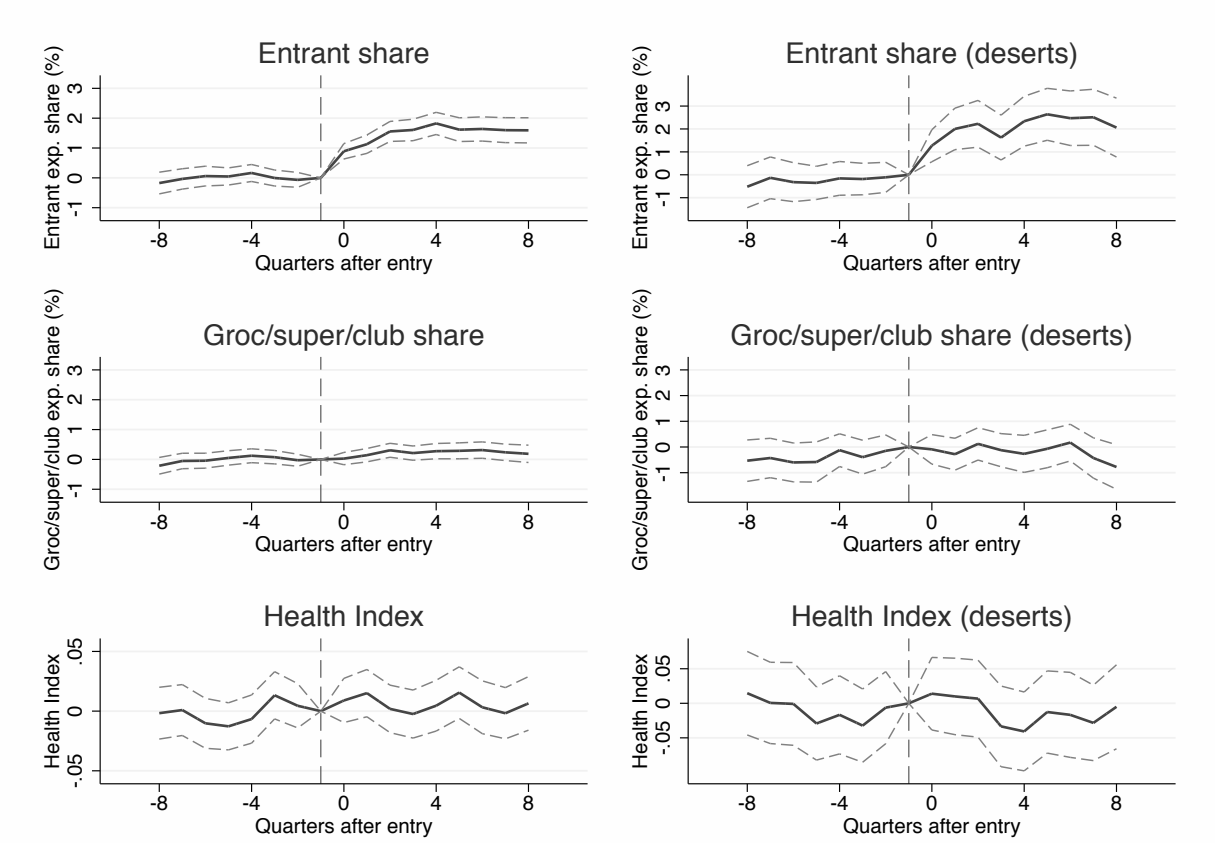
C.A Additional Figures and Tables for Entry Event Study

Figure A4: **Event Study of Supermarket Entry Between 10 and 15 Minutes from Home**



Notes: This figure presents the $\tau_{[10,15)q}$ parameters and 95 percent confidence intervals from estimates of Equation (2): the effects of supermarket entry, using 2004–2016 household-by-quarter Homescan data. All regressions control for household demographics (natural log of income, natural log of years of education, age indicators, household size, race indicators, a married indicator, employment status, and weekly work hours), census division-by-quarter of sample indicators, and household-by-census tract fixed effects. The top two panels present effects on expenditure shares (the share of all grocery expenditures recorded in Homescan, in units of percentage points) across all retailers with stores that have entered within a 15-minute drive of the household. The middle two panels present effects on the combined expenditure share of grocery stores, supercenters, and club stores. We keep the y-axis on the same scale between the top and middle panels so that the magnitudes can be easily compared. The bottom two panels present effects on the Health Index, our overall measure of the healthfulness of grocery purchases which is normalized to have a mean of zero and a standard deviation of one across households. The left panels include the full sample, while the right panels include only the "food desert" subsample: observations with no grocery stores with 50 or more employees, supercenters, or club stores in the zip code in the first year the household is observed there. Observations are not weighted for national representativeness.

Figure A5: **Event Study of Supermarket Entry with Additional Leads**



Notes: This figure presents the $\tau_{[0,10)q}$ parameters and 95 percent confidence intervals from estimates of Equation (2): the effects of supermarket entry, using 2004–2016 household-by-quarter Homescan data. The figure parallels Figure 4, except with additional leads of the entry. All regressions control for household demographics (natural log of income, natural log of years of education, age indicators, household size, race indicators, a married indicator, employment status, and weekly work hours), census division-by-quarter of sample indicators, and household-by-census tract fixed effects. The top two panels present effects on expenditure shares (the share of all grocery expenditures recorded in Homescan, in units of percentage points) across all retailers with stores that have entered within a 15-minute drive of the household. The middle two panels present effects on the combined expenditure share of grocery stores, supercenters, and club stores. We keep the y-axis on the same scale between the top and middle panels so that the magnitudes can be easily compared. The bottom two panels present effects on the Health Index, our overall measure of the healthfulness of grocery purchases which is normalized to have a mean zero and a standard deviation of one across households. The left panels include the full sample, while the right panels include only the "food desert" subsample: observations with no grocery stores with 50 or more employees, supercenters, or club stores in the zip code in the first year the household is observed there. Observations are not weighted for national representativeness.

Table A4: **Effects of Supermarket Entry**

(a) **Effects on Expenditure Shares at Drug and Convenience Stores**

	Full sample	Bottom quartile	Food deserts
	(1)	(2)	(3)
Post entry: 0-10 minutes	-0.015	-0.048	0.058
	(0.022)	(0.065)	(0.067)
Post entry: 10-15 minutes	-0.033*	-0.053	-0.112*
	(0.017)	(0.047)	(0.064)
Observations	2,874,365	537,998	646,181
Dependent var. mean	2.6	3.1	2.4

(b) **Effects on Health Index Using Alternative Food Desert Definitions**

	<1000 produce UPCs	No medium groceries	Three-mile radius
	(1)	(2)	(3)
Post entry: 0-10 minutes	0.004	0.013	0.017
	(0.011)	(0.012)	(0.014)
Post entry: 10-15 minutes	0.006	0.004	0.019*
	(0.007)	(0.007)	(0.010)
Observations	411,654	378,682	490,551

Notes: This table uses 2004–2016 Nielsen Homescan data at the household-by-quarter level. The table parallels Table 2, except Panel (a) presents effects on expenditure shares at drug and convenience stores, and Panel (b) uses alternative definitions of a "food desert." In Panel (b), column 1 defines food deserts as zip codes with fewer than 1,000 produce UPCs, as predicted by projecting produce UPC counts in RMS stores from column 3 of Table A3 onto Zip Code Business Patterns store count data; column 2 uses the primary food desert definition but also excludes any zip codes with grocery stores employing between 10 and 49 employees; and column 3 defines a zip code as a food desert only if all zip codes with centroids within three miles have no grocery stores with 50 or more employees, supercenters, or club stores. Expenditure shares are the share of total grocery expenditures recorded in Homescan, in units of percentage points. The Health Index is our overall measure of the healthfulness of grocery purchases and is normalized to have a mean of zero and a standard deviation of one across households. Reported independent variables are the count of supermarkets that have entered within a 0–10 or 10–15 minute drive from the household's census tract centroid. All regressions control for household demographics (natural log of income, natural log of years of education, age indicators, household size, race indicators, a married indicator, employment status, and weekly work hours), census division-by-quarter of sample indicators, and household-by-census tract fixed effects. Observations are not weighted for national representativeness. Robust standard errors, clustered by household and census tract, are in parentheses. *, **, ***: statistically significant with 10, 5, and 1 percent confidence, respectively.

Table A5: **Effects of Supercenter Entry**

(a) **Effects on Expenditure Shares**

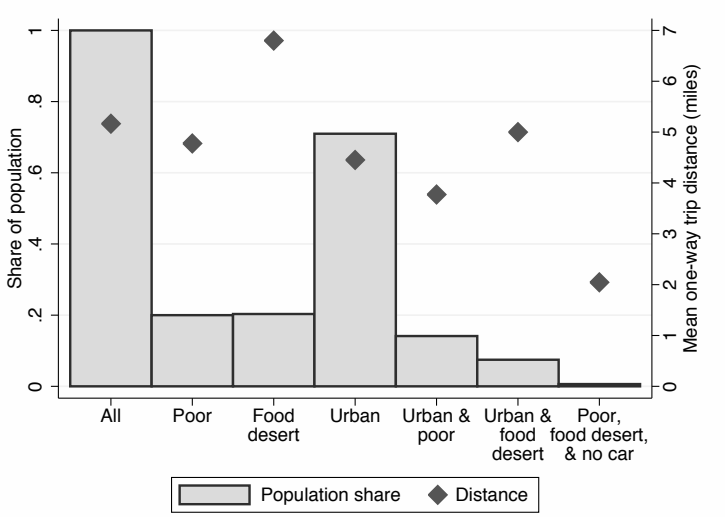
	Full sample		Bottom quartile		Food deserts	
	(1)	(2)	(3)	(4)	(5)	(6)
	Entrants	Grocery/ super/club	Entrants	Grocery/ super/club	Entrants	Grocery/ super/club
Post entry: 0-10 minutes	3.786***	0.439***	5.224***	0.535*	3.629***	0.227
	(0.226)	(0.107)	(0.629)	(0.314)	(0.624)	(0.302)
Post entry: 10-15 minutes	1.601***	0.033	1.902***	0.238	1.976***	0.205
	(0.114)	(0.071)	(0.345)	(0.209)	(0.292)	(0.183)
Observations	2,874,514	2,874,365	538,041	537,998	646,223	646,181
Dependent var. mean	3.7	88.2	3.4	86.2	2.6	87.7

(b) **Effects on Health Index**

	Full sample	Bottom quartile	Food deserts
	(1)	(2)	(3)
Post entry: 0-10 minutes	0.008	0.019	0.018
	(0.006)	(0.015)	(0.016)
Post entry: 10-15 minutes	0.006	-0.002	0.020*
	(0.004)	(0.011)	(0.011)
Observations	2,874,514	538,041	646,223

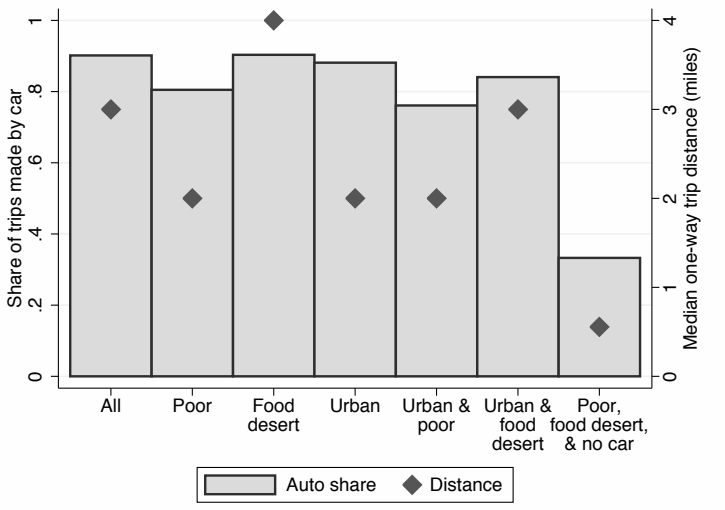
Notes: This table uses 2004-2016 Nielsen Homescan data at the household-by-quarter level. It parallels Table 2, except it considers entry by supercenters only, excluding other types of grocery stores. The "food desert" subsample comprises observations with no grocery stores with 50 or more employees, supercenters, or club stores in the zip code in the first year the household is observed there. Expenditure shares are the share of total grocery expenditures recorded in Homescan, in units of percentage points. The Health Index is our overall measure of the healthfulness of grocery purchases and is normalized to have a mean of zero and a standard deviation of one across households. Reported independent variables are the count of supermarkets that have entered within a 0–10 or 10–15 minute drive from the household's census tract centroid. All regressions control for household demographics (natural log of income, natural log of years of education, age indicators, household size, race indicators, a married indicator, employment status, and weekly work hours), census division-by-quarter of sample indicators, and household-by-census tract fixed effects. Observations are not weighted for national representativeness. Robust standard errors, clustered by household and census tract, are in parentheses. *, **, ***: statistically significant with 10, 5, and 1 percent confidence, respectively.

Figure A6: **Shopping Trip Distances by Household Income**



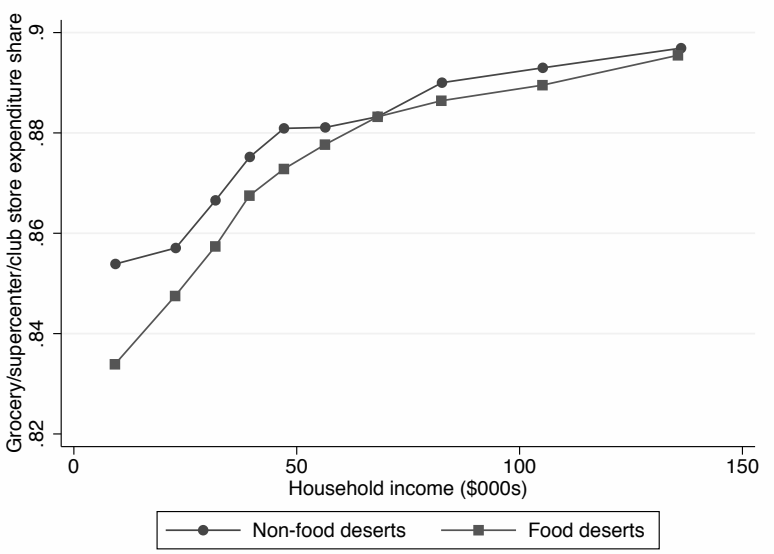
Notes: Data are from the 2009 National Household Travel Survey. Diamonds represent the mean one-way trip distance for trips beginning or ending in "buying goods: groceries/clothing/hardware store." "Poor" means households in the bottom income quartile. "Food desert" means that the household is in a zip code with no grocery stores with 50 or more employees, supercenters, or club stores. "Urban" includes urbanized areas or urban clusters of at least 2500 people, using the U.S. Census Bureau definition. "No car" means that the household does not own a car.

Figure A7: **Median Shopping Trip Distances by Household Income**



Notes: Data are from the 2009 National Household Travel Survey. Diamonds represent the median one-way trip distance for trips beginning or ending in "buying goods: groceries/clothing/hardware store." "Poor" means households in the bottom income quartile. "Food desert" means that the household is in a zip code with no grocery stores with 50 or more employees, supercenters, or club stores. "Urban" includes urbanized areas or urban clusters of at least 2500 people, using the U.S. Census Bureau definition. "No car" means that the household does not own a car.

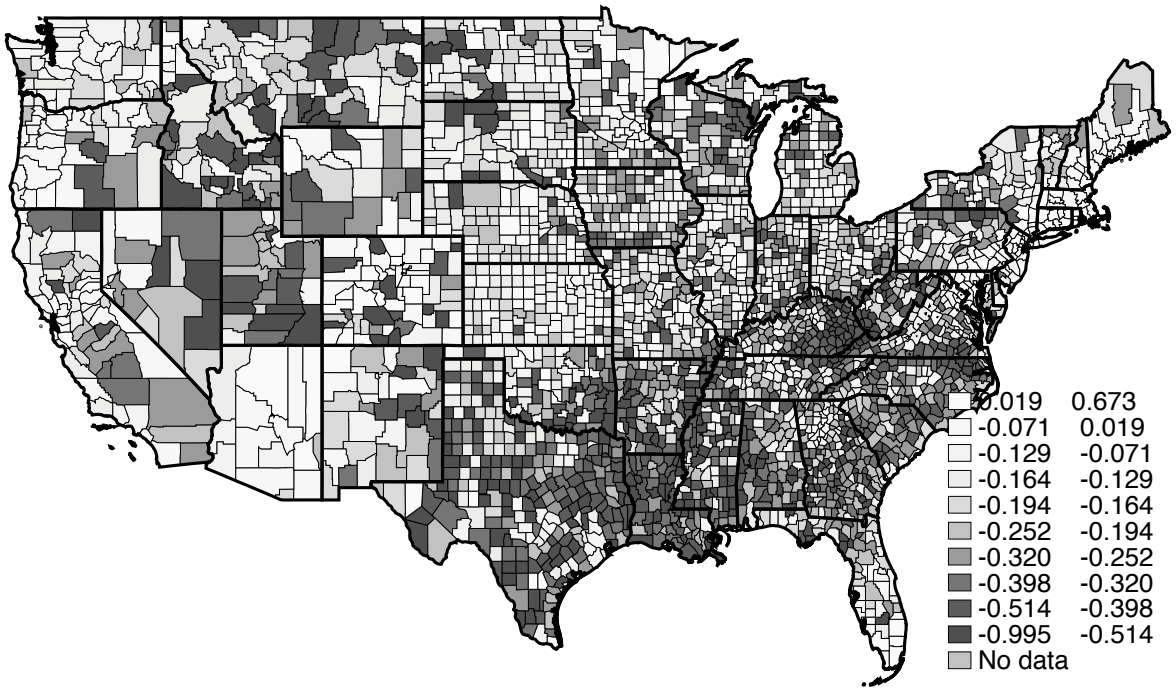
Figure A8: **Supermarket Expenditure Shares by Household Income**



Notes: This figure presents the share of grocery expenditures that are made at grocery stores, supercenters, and club stores against average household income across all years the household is observed in Homescan, residual of age and year indicators and household size, using Nielsen Homescan data for 2004–2016. A household-by-year observation is in a "food desert" if its zip code does not have any grocery stores with 50 or more employees, supercenters, or club stores in that year. Observations are weighted for national representativeness.

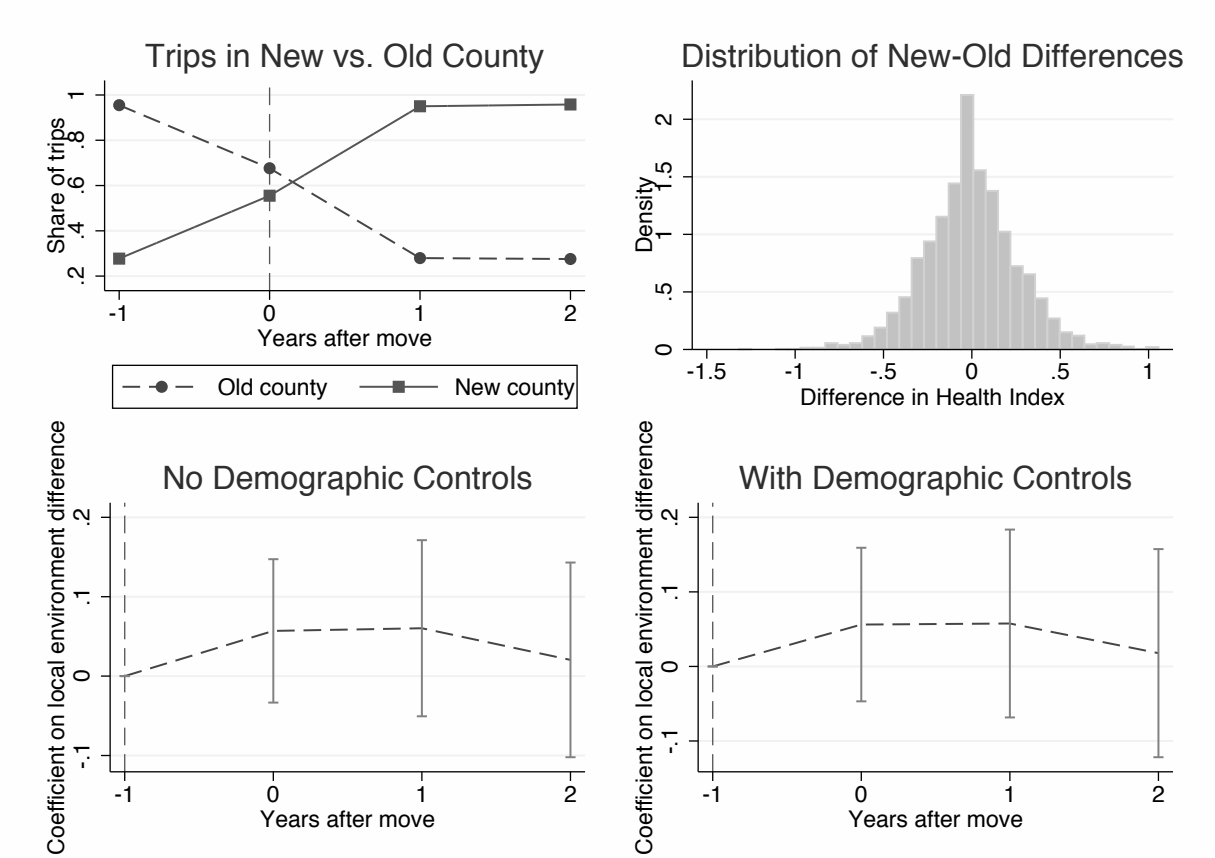
C.B Appendix to Movers Event Study

Figure A9: **Average Health Index of Store Purchases by County**



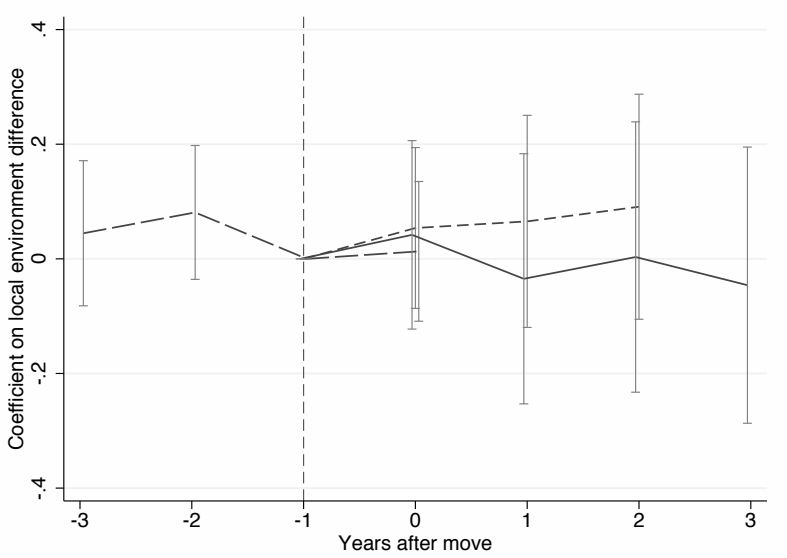
Notes: This figure presents the calorie-weighted average normalized Health Index of packaged grocery purchases by county, using 2006–2016 Nielsen RMS data. The Health Index is our overall measure of the healthfulness of grocery purchases and is normalized to have a mean of zero and a standard deviation of one across households. Because the RMS data do not contain the complete census of stores, the distribution of store types in the RMS sample may not match a county's true distribution. For example, the RMS sample might include most of the grocery stores in county A, but few of the grocery stores and most of the drug stores in county B. To estimate the county average Health Index, we thus take the calorie-weighted average Health Index of groceries sold in RMS stores and regression-adjust for the difference between the distribution of store channel types in the RMS data versus the true distribution of store channel types observed in ZBP data. Note that purchases in RMS are less healthful than in Homescan, so the average normalized Health Index on this map is less than zero.

Figure A10: **Event Study of Moves Across Zip Codes**

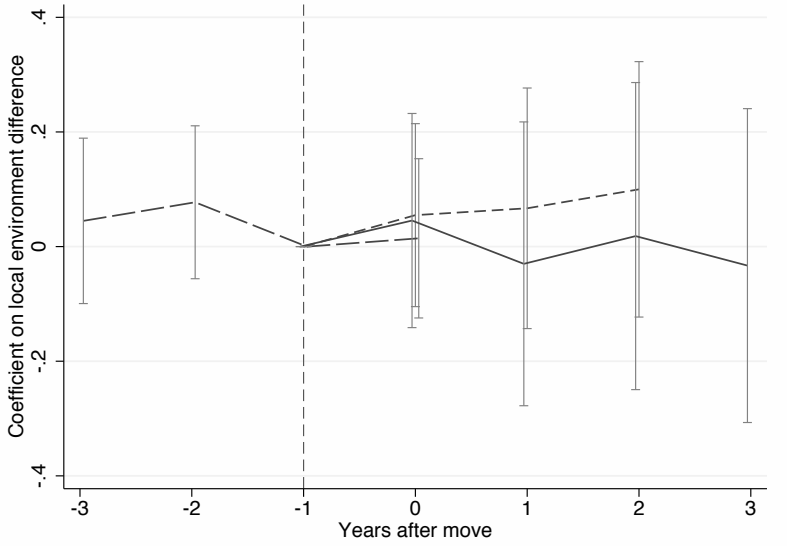


Notes: Using 2004–2016 Homescan data, these figures present results for the event study of moves across zip codes. The top left panel presents the share of shopping trips that are in the new versus old county. The top right panel presents the distribution across balanced panel households of the difference in the Health Index between the new and old zip code. The bottom panels present the τ_y parameters and 95 percent confidence intervals from estimates of Equation (4): associations between the household-level Health Index and the difference in the average local Health Index between post-move and pre-move locations. The bottom right panel includes controls for household demographics (natural log of income, natural log of years of education, age indicators, household size, race indicators, a married indicator, employment status, and weekly work hours). The Health Index is our overall measure of the healthfulness of grocery purchases and is normalized to have a mean of zero and a standard deviation of one across households. Observations are not weighted for national representativeness.

Figure A11: **Event Study of Movers with Different Balanced Panel Windows**



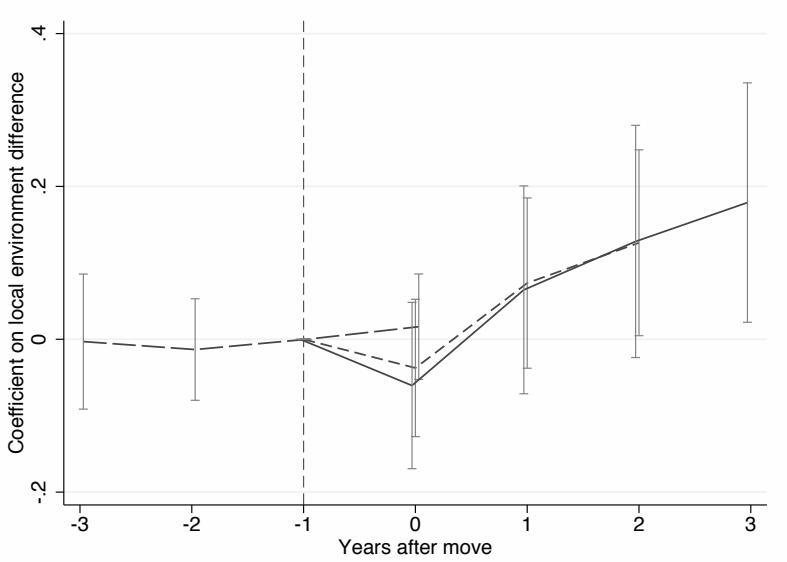
(a) **Without Controls**



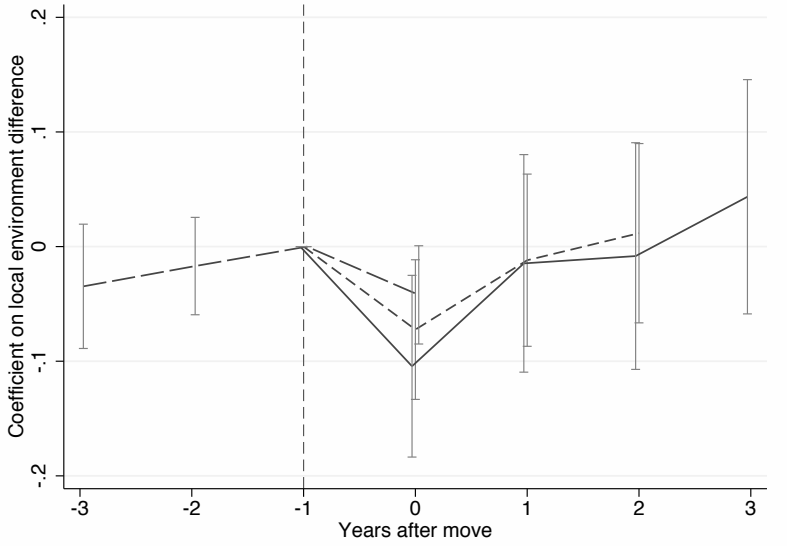
(b) **With Controls**

Notes: Using 2004–2016 Homescan data, these figures present the τ_y parameters and 95 percent confidence intervals from estimates of Equation (4): associations between the household-level Health Index and the difference in the average local Health Index between post-move and pre-move locations. Each figure superimposes three different estimates identified off of balanced panels for different windows around the move. Panel (b) includes controls for household demographics (natural log of income, natural log of years of education, age indicators, household size, race indicators, a married indicator, employment status, and weekly work hours). The Health Index is our overall measure of the healthfulness of grocery purchases and is normalized to have a mean of zero and a standard deviation of one across households. Observations are not weighted for national representativeness.

Figure A12: **Event Study: Income Changes in Mover Households**



(a) **Moves Across Counties**



(b) **Moves Across Zip Codes**

Notes: Using 2004–2016 Homescan data, these figures present the τ_y parameters and 95 percent confidence intervals from estimates of Equation (4): associations between natural log of household income and the difference in the average local Health Index between post-move and pre-move locations. All regressions control for year indicators and household fixed effects. Each figure superimposes three different estimates identified off of balanced panels for different windows around the move. The Health Index is our overall measure of the healthfulness of grocery purchases and is normalized to have a mean of zero and a standard deviation of one across households. Observations are not weighted for national representativeness. The regressions are the same as in Figure 5, except with natural log of household income as the dependent variable and no controls for household demographics.

Table A6: **Association of Income with Local Area Health Index Using Movers**

	(1)	(2)
Zip code average Health Index	0.00590	
	(0.36)	
County average Health Index		0.125***
		(3.25)
Observations	564,944	570,279
95% confidence interval upper bound	0.038	0.200

Notes: This table uses 2004–2016 Nielsen Homescan data at the household-by-year level. The sample excludes observations where less than 50 percent of trips to RMS stores are not in the household's end-of-year county of residence. The dependent variable is the natural log of household income. The Health Index is our overall measure of the healthfulness of grocery purchases and is normalized to have a mean of zero and a standard deviation of one across households. All regressions control for year indicators and household fixed effects. Observations are not weighted for national representativeness. Robust standard errors, clustered by household and local area (zip code or county), are in parentheses. *, **, ***: statistically significant with 10, 5, and 1 percent confidence, respectively.

Table A7: **Association of Coke Market Share with Local Area Coke Market Share Using Movers**

	(1)	(2)
County average Coke market share	0.1620***	0.1613***
	(0.0418)	(0.0417)
Household demographics	No	Yes
Observations	306,714	306,714

Notes: This table uses 2004–2016 Nielsen Homescan data at the household-by-year level. The sample excludes observations where less than 50 percent of trips to RMS stores are not in the household's end-of-year county of residence. Coke market share equals Coke calories purchased / (Coke + Pepsi calories purchased). Household demographics are natural log of income, natural log of years of education, age indicators, household size, race indicators, a married indicator, employment status, and weekly work hours. All regressions also control for year indicators and household fixed effects. Observations are not weighted for national representativeness. Robust standard errors, clustered by household and county, are in parentheses. *, **, ***: statistically significant with 10, 5, and 1 percent confidence, respectively.

D Appendix to Demand Model Estimation

D.A Derivation of Annual Calorie Demand

In this appendix, we show that we can aggregate the first-order conditions from Equation (6) over time to obtain annual calorie demand in Equation (7). Let τ index subperiods (such as weeks or shopping trips) within year t. Since the first-order conditions from Equation (6) hold in each subperiod, we can aggregate these first-order conditions over τ to the household-by-year level:

$$\sum_{\tau \epsilon t} \sum_{k=1}^{K_j} p_{kj\tau} y_{ikj\tau} = \sum_{c=1}^{C} \frac{\beta_c}{\lambda_i} \sum_{k=1}^{K_j} \sum_{\tau \epsilon t} a_{kjc} y_{ikj\tau} + \sum_{\tau \epsilon t} \frac{\mu_{ij\tau} \theta_{ij\tau}}{\lambda_i}. \tag{A1}$$

Equation (A1) illustrates one feature of the model: it allows for estimation of characteristic and product group preferences from data aggregated to the level of household-by-product group-by-year. While the model is precisely microfounded, this aggregation allows us to avoid dealing with parameters driving UPC-level preferences and weekly dynamics such as stockpiling. The model potentially allows for considerable heterogeneity across households and time. As described below, we will estimate separate parameters for each of four household income groups, assuming homogeneous β_c parameters within each group.

To economize on notation, define total calories purchased by household i in product group j in year t as $Y_{ijt} = \sum_{\tau \epsilon t} \sum_{k \epsilon K_j} y_{ikj\tau}$. Define $\tilde{p}_{ijt}$ and $\tilde{a}_{ijct}$, respectively, as the average price paid per calorie and average amount of characteristic c per calorie for household i's purchases in group j in year t. Define $\tilde{\delta}_{ijt} = \sum_{\tau \epsilon t} \frac{\mu_{ij\tau} \theta_{ij\tau}}{\lambda_i}$ as the strength of household i's preferences for group j in year t. Finally, let $\tilde{\beta}_c = \frac{\beta_c}{\lambda_i}$ be the money-metric marginal utility of each characteristic. Equation (A1) can now be written more compactly as:

$$\tilde{p}_{ijt} Y_{ijt} = \sum_{c=1}^{C} \tilde{\beta}_c \tilde{a}_{ijct} Y_{ijt} + \tilde{\delta}_{ijt}. \tag{A2}$$

As in the existing literature that uses the characteristics approach to model demand, we allow for a product characteristic that is unobserved to the econometrician (Berry, 1994). Let characteristic $c = 1$ be unobserved, and let characteristics $c = 2, ..., C$ be observed. We denote $\xi = \tilde{\beta}_1 \tilde{a}_{ij1t}$ as the unobserved characteristic, which again is assumed to be constant within each income group.

We now depart from the empirical strategy in Dubois, Griffith and Nevo (2014).[1] Instead of estimating Equation (A2) directly, we solve for total calories Y_{ijt} and take logs of both sides. We

[1] Dubois, Griffith and Nevo (2014) directly estimate the average nutrient preference parameters β_c using a version of Equation (A2):

$$\tilde{p}_{ijt} Y_{ijt} = \sum_{c=2}^{C} \tilde{\beta}_c \tilde{a}_{ijct} Y_{ijt} + \epsilon_{ijt}, \tag{A3}$$

including additional fixed effects and instrumenting for $\tilde{a}_{ijct} Y_{ijt}$ using variation in local product availability. Unfortunately, the error term in this regression contains purchases of the unobserved nutrient: $\epsilon_{ijt} = \xi Y_{ijt} + \tilde{\delta}_{ijt}$. Thus, if the instrument affects consumption Y_{ijt}, it *mechanically* is also correlated with the error term. No instrument can address this mechanical endogeneity problem.

also separate the household's product group preferences $\ln \tilde{\delta}_{ijt}$ into a product group fixed effect δ_j, a geographic market fixed effect ϕ_m (which in practice will be a county), a year fixed effect ϕ_t, and the household-specific deviation ε_{ijt}, so $\ln \tilde{\delta}_{ijt} = \delta_j + \phi_m + \phi_t + \varepsilon_{ijt}$. Our final estimating equation for each income group is thus

$$\ln Y_{ijt} = -\ln\left(\tilde{p}_{ijt} - \sum_{c=2}^{C} \tilde{\beta}_c \tilde{a}_{ijct} - \xi\right) + \delta_j + \phi_m + \phi_t + \varepsilon_{ijt}. \tag{A4}$$

D.B Method of Moments Estimation

Our method of moments estimator is defined as follows:

$$\left(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\phi}}, \hat{\tilde{\boldsymbol{\beta}}}, \hat{\xi}\right) = \underset{(\boldsymbol{\delta},\boldsymbol{\phi},\tilde{\boldsymbol{\beta}},\xi)}{\arg\min} \left(\frac{1}{IJT}\sum_i\sum_j\sum_t \boldsymbol{g}_{ijt}\right)' \left(\frac{1}{IJT}\sum_i\sum_j\sum_t \boldsymbol{g}_{ijt}\right). \tag{A5}$$

Define $\boldsymbol{Y}$ as the vector of product group calorie consumption Y_{ijt}, $\boldsymbol{F}\left(\tilde{\boldsymbol{\beta}}, \xi\right)$ as the vector of implicit prices $F_{ijt} = \left(\tilde{p}_{ijt} - \sum_{c=2}^{C} \tilde{\beta}_c \tilde{a}_{ijct} - \xi\right)$, $\boldsymbol{D}$ as a stacked matrix of the two dummy variable matrices ($\boldsymbol{D}_j$ and $\boldsymbol{D}_m$), $\boldsymbol{Z}$ as a matrix with all of our vectors of instruments ($\boldsymbol{D}$, the nutrient content $\tilde{\boldsymbol{a}}$, and the price instruments $\boldsymbol{P}$), and $\boldsymbol{Pr_D} = \left(\boldsymbol{D'ZZ'D}\right)^{-1}\boldsymbol{D'ZZ'}$ as a projection matrix. We can simplify the estimation problem by solving for our vectors of linear coefficients, $\boldsymbol{\delta}$ and $\boldsymbol{\phi}$, as analytic functions of $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\xi}$:

$$(\boldsymbol{\delta}, \boldsymbol{\phi}) = \boldsymbol{Pr}_D\left(\ln(\boldsymbol{Y}) - \boldsymbol{F}\left(\tilde{\boldsymbol{\beta}}, \xi\right)\right). \tag{A6}$$

Substituting Equation (A6) back into Equation (A5), we can re-write the MOM estimator in terms of $\tilde{\boldsymbol{\beta}}$ and $\hat{\xi}$:

$$\left(\hat{\tilde{\boldsymbol{\beta}}}, \hat{\xi}\right) = \underset{(\tilde{\boldsymbol{\beta}},\xi)}{\arg\min} \left(\frac{1}{IJT}\sum_i\sum_j\sum_t g_{ijt}\left(\tilde{\boldsymbol{\beta}}, \xi\right)\right)' \left(\frac{1}{IJT}\sum_i\sum_j\sum_t g_{ijt}\left(\tilde{\boldsymbol{\beta}}, \xi\right)\right). \tag{A7}$$

At the true value, the gradient for this problem is:

$$-2\boldsymbol{G}\left(\tilde{\boldsymbol{\beta}}, \xi\right)'\boldsymbol{G}\left(\tilde{\boldsymbol{\beta}}, \xi\right) = 0 \tag{A8}$$

where the Jacobian of the moments, $\boldsymbol{G}\left(\tilde{\boldsymbol{\beta}}, \xi\right)$, is

$$\boldsymbol{G}\left(\tilde{\boldsymbol{\beta}}, \xi\right) = \frac{1}{IJT}\begin{bmatrix} \tilde{\boldsymbol{a}}'\left(\boldsymbol{I} - \boldsymbol{DPr_D}\right) \\ \boldsymbol{P}'\left(\boldsymbol{I} - \boldsymbol{D}_m\boldsymbol{Pr}_{\boldsymbol{D}_m}\right) \\ \boldsymbol{D}'\left(\boldsymbol{I} - \boldsymbol{DPr_D}\right) \end{bmatrix} \nabla_\beta \boldsymbol{F}\left(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{a}}; \tilde{\boldsymbol{\beta}}\right). \tag{A9}$$

In the above equation, $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{Pr}_{\boldsymbol{D}_m}$ is a projection matrix using $\boldsymbol{D}_m$.

The covariance matrix of our full MOM estimator, $\boldsymbol{\Theta}^{MOM} \equiv \left(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\phi}}, \hat{\bar{\beta}}\hat{\xi}\right)$, is $cov\left(\boldsymbol{\Theta}^{MOM}\right) = \left(\boldsymbol{G}'\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\boldsymbol{\Omega}\boldsymbol{G}\left(\boldsymbol{G}'\boldsymbol{G}\right)^{-1}$, with Jacobian matrix

$$G = \frac{1}{IJT}\sum_i\sum_j\sum_t \begin{bmatrix} \overrightarrow{0}'_J & -\tilde{a}_{ijt}D'_m & -\tilde{a}_{ijt}\nabla_\beta F'_{ijt} \\ -P_{jmt}D'_j & -P_{jmt}D'_m & -P_{jmt}\nabla_\beta F'_{ijt} \\ -D_jD'_j & -D_jD'_m & -D_j\nabla_\beta F'_{ijt} \\ -D_mD'_{ijt} & -D_mD'_m & -D_m\nabla_\beta F'_{ijt} \end{bmatrix} \tag{A10}$$

and covariance matrix

$$\boldsymbol{\Omega} \;=\; \mathbb{E}\left(\boldsymbol{g}_{ijt}\left(\boldsymbol{\Theta}^{MOM}\right)\boldsymbol{g}_{ijt}\left(\boldsymbol{\Theta}^{MOM}\right)'\right). \tag{A11}$$

When computing our standard errors, we cluster by household as follows:

$$\hat{\Omega} = \frac{1}{IJT}\sum_i\sum_{j,j'}\sum_{t,t'}\boldsymbol{g}_{ijt}\left(\hat{\bar{\boldsymbol{\beta}}},\hat{\xi}\right)\boldsymbol{g}_{ij't'}\left(\hat{\bar{\boldsymbol{\beta}}},\hat{\xi}\right)'. \tag{A12}$$

D.C Describing the Instrument

Table A8: **Source of Variation in the Instrument**

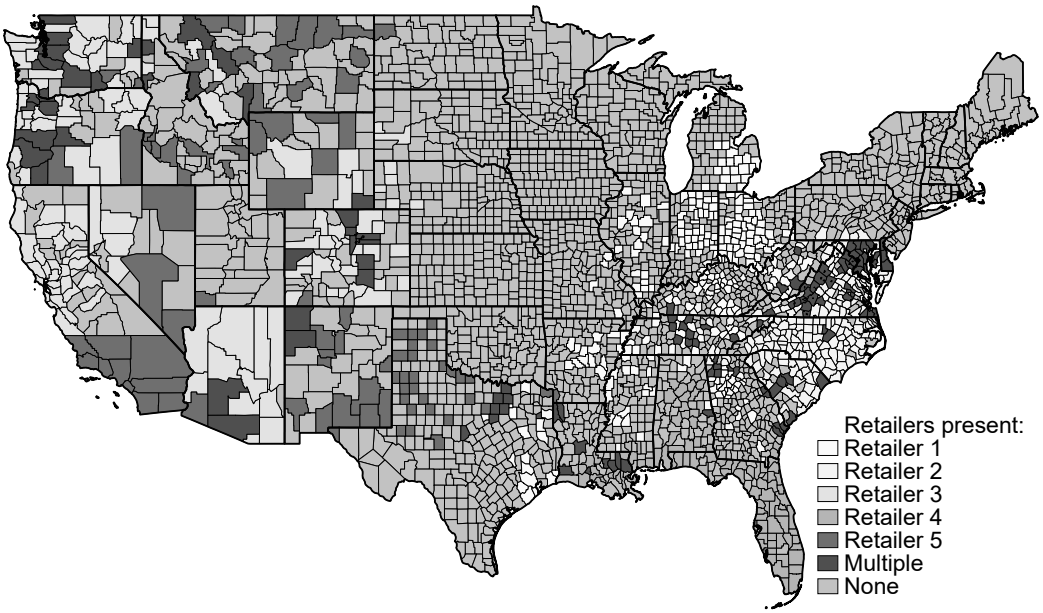
	(1)	(2)	(3)	(4)	(5)	(6)
Unadjusted R^2	0.426	0.441	0.483	0.699	0.452	0.879
Adjusted R^2	0.426	0.441	0.469	0.699	0.452	0.877
County, product group, retailer fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
State-product group fixed effects	No	Yes	No	No	No	No
County-product group fixed effects	No	No	Yes	No	No	No
Retailer-product group fixed effects	No	No	No	Yes	No	No
Year-product group fixed effects	No	No	No	No	Yes	No
Retailer-year-product group fixed effects	No	No	No	No	No	Yes

Notes: Let P_{rjmt} be retailer r's cost advantage in product group j:

$$P_{rjmt} = \frac{\sum_{k=1}^{\mathcal{K}_j} N_{kt}\Delta\ln(p_{krt,-m})}{\sum_{k=1}^{\mathcal{K}_j} N_{kt}}. \tag{A13}$$

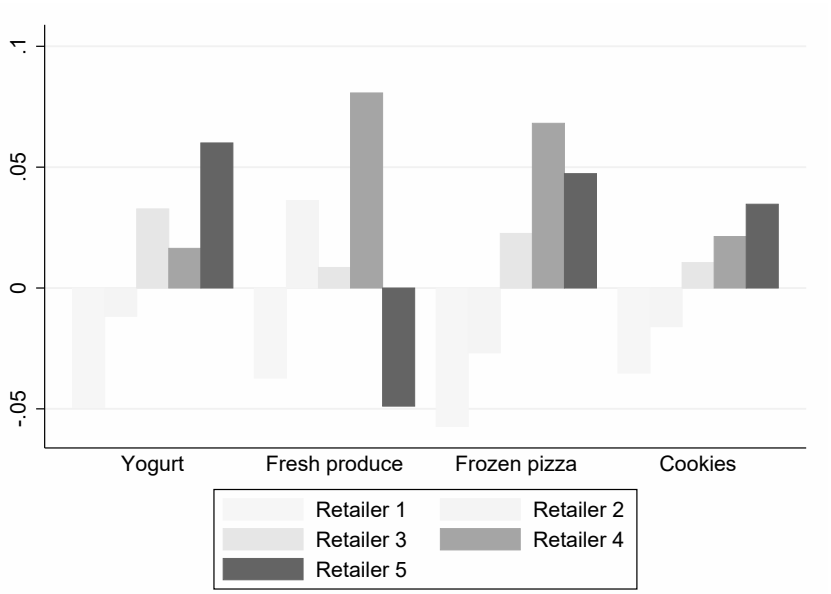
In other words, P_{rjmt} is the instrument at the chain-by-county level, before averaging across chains to the county level. This table presents the R^2 of regressions of P_{rjmt} on different vectors of fixed effects, thereby illustrating the sources of variation in the instrument.

Figure A13: **Geographic Variation in Presence of Large Retail Chains**



Notes: This figure shows the counties where the largest five large retail chains in RMS had stores in 2015.

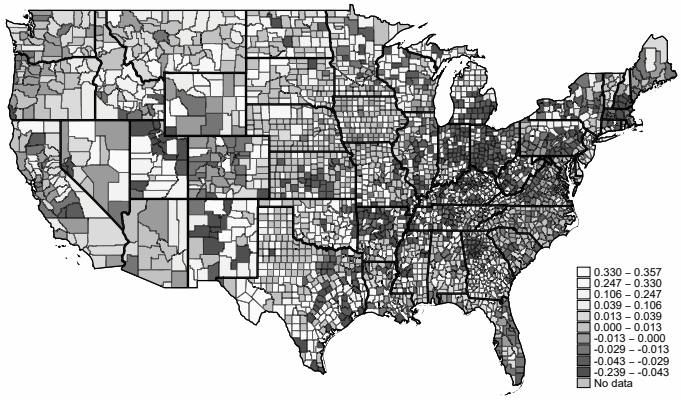
Figure A14: **Price Variation for Large Retail Chains**



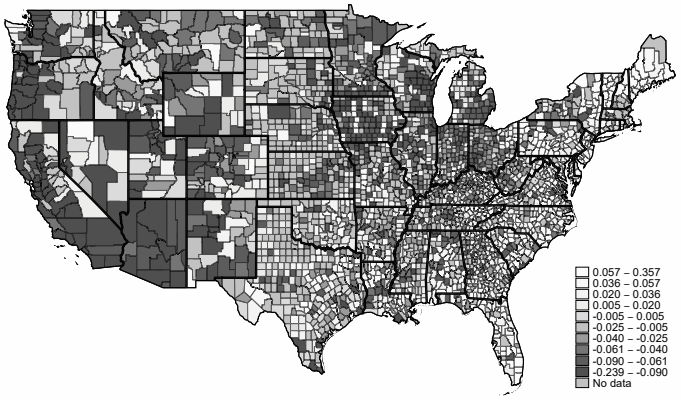
Notes: Let P_{rjmt} be retailer r's cost advantage in product group j, as defined in Equation (A13). This figure shows the average value of P_{rjmt} for the largest five retail chains in RMS for four example product groups.

Figure A15: **Geographic Variation in the Price Instrument**

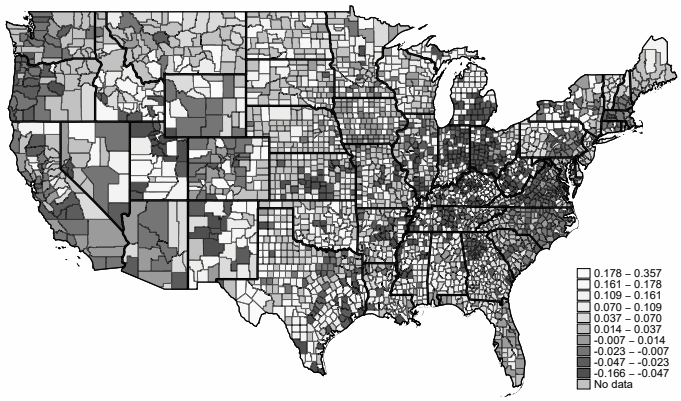
(a) **Yogurt**



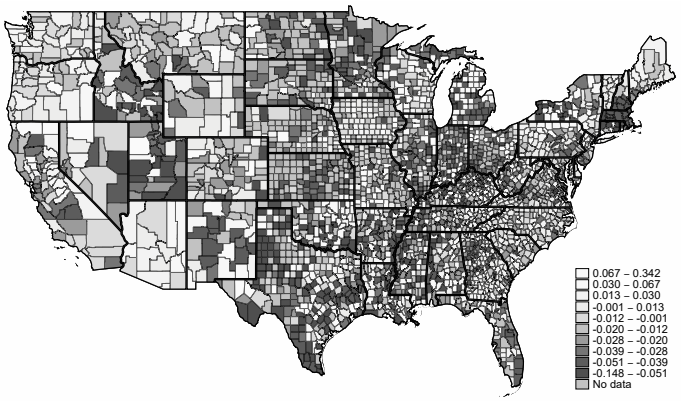
(b) **Fresh Produce**



(c) **Frozen Pizza**

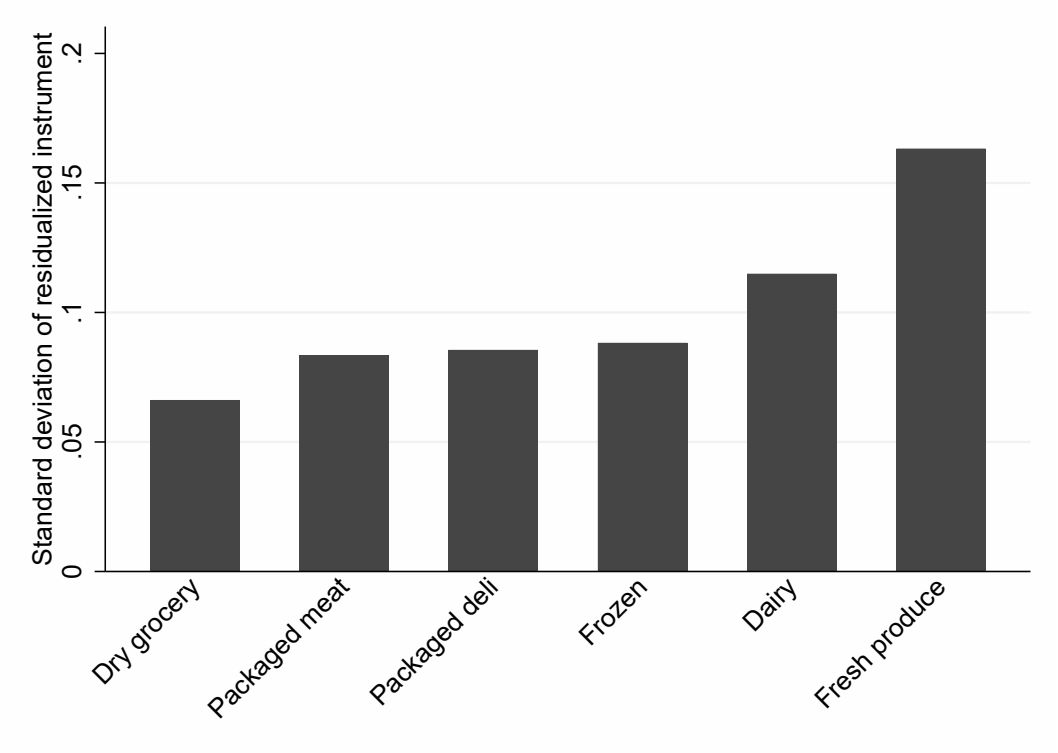


(d) **Cookies**



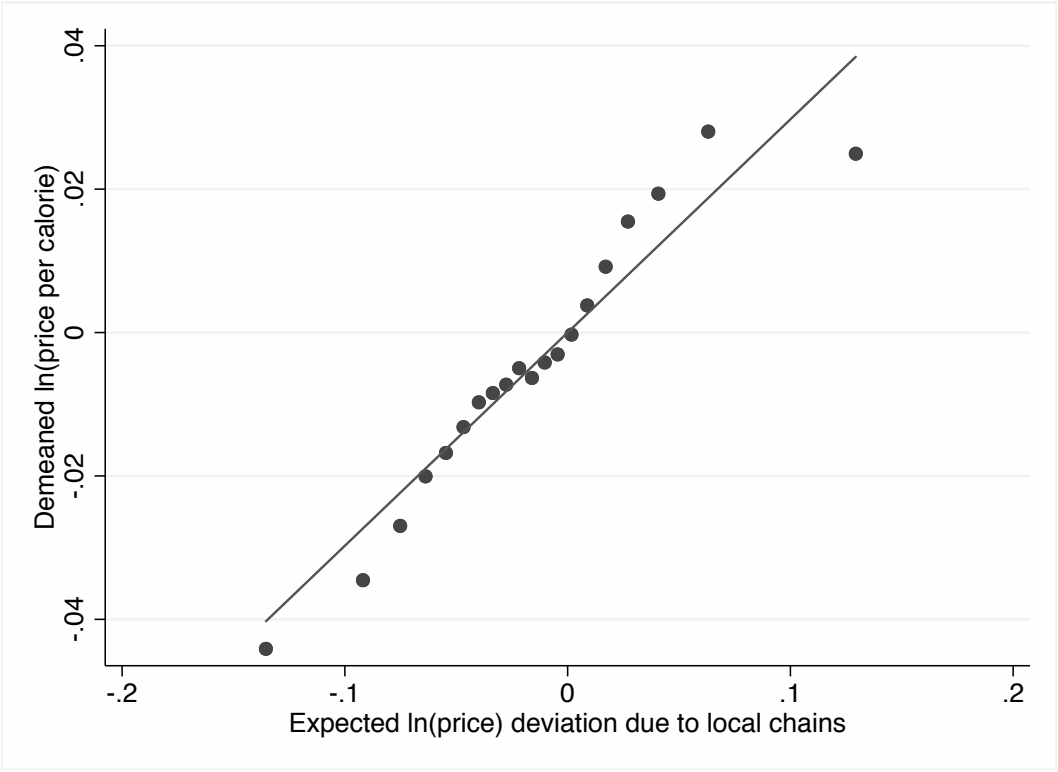
Notes: These figures present the county averages (over the years of our sample) of the price instrument P_{jmt} for four example product groups.

Figure A16: **Standard Deviation of Price Instrument by Product Department**



Notes: This figure presents the standard deviation of our price instrument P_{jmt}, after residualizing against year, product group, and county fixed effects. The instrument is in units of log price per calorie.

Figure A17: **Binned Scatterplot of First Stage Price Regression**



Notes: This figure presents a binned scatterplot of a regression of natural log price per calorie on our price instrument P_{jmt}, residual of product group-by-income quartile, county-by-income quartile, and year-by-income quartile fixed effects.

Table A9: **Suggestive Tests of the Exclusion Restriction**

	(1)	(2)
	Demand predicted by demographics	Health Index × ln(County median income)
Price instrument	-0.006	-0.044
	(0.007)	(0.028)

Notes: This table presents suggestive tests of the exclusion restriction using the following regression:

$$\hat{Y}_{jmt} = \pi P_{jmt} + \delta_j + \phi_m + \phi_t + \varepsilon_{jmt}, \tag{A14}$$

where $\hat{Y}_{jmt}$ are predictors of demand for groceries in product group j in market m in year t. For column 1, we first estimate the relationship between purchases of product group j and household demographics using the following regression: $\ln Y_{ijt} = \gamma_j \boldsymbol{X}_{it} + \delta_{jmt} + \varepsilon_{ijt}$,where $\boldsymbol{X}_{it}$ is the first seven household covariates presented in Table 1 and δ_{jmt} are product group-by-county-by year fixed effects, included to ensure that the estimates capture variation in preferences across households within the same market. We then predict county average purchases as $\hat{Y}_{jmt} = \hat{\gamma}_j \bar{\boldsymbol{X}}_{mt}$, where $\bar{\boldsymbol{X}}_{mt}$ is county average demographics. For column 2, we let $\hat{Y}_{jmt}$ equal the interaction of the average Health Index across UPCs in product group j and the natural log of county m's median income in year t. Standard errors, clustered by county, are in parentheses. *, **, ***: statistically significant with 10, 5, and 1 percent confidence, respectively.

D.D Additional Tables and Figures

Table A10: **Preferences for Nutrients by Household Income**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Income quartile	Sodium	Whole fruit	Other fruit	Whole grains	Refined grains	Greens, beans	Other veg	Dairy
Income Q1	-0.178***	-0.324***	-0.064***	0.177***	-0.014***	0.840***	-0.290***	0.084***
	(0.009)	(0.020)	(0.008)	(0.003)	(0.0004)	(0.007)	(0.016)	(0.003)
Income Q2	-0.299***	-0.225***	-0.086***	0.228***	-0.010***	0.861***	-0.382***	0.063***
	(0.012)	(0.015)	(0.007)	(0.005)	(0.0005)	(0.007)	(0.017)	(0.003)
Income Q3	-0.384***	-0.233***	-0.093***	0.269***	-0.0096***	0.970***	-0.436***	0.068***
	(0.014)	(0.015)	(0.007)	(0.006)	(0.001)	(0.009)	(0.019)	(0.003)
Income Q4	-0.585***	-0.246***	-0.115***	0.365***	0.0031*	1.216***	-0.605***	0.084***
	(0.025)	(0.020)	(0.010)	(0.011)	(0.0017)	(0.018)	(0.031)	(0.004)
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Sea, plant protein	Meat protein	Added sugar	Solid fats	Unobserved characteristic	Shelf life	Convenience	WTP for Health Index
Income Q1	-0.240***	0.027***	0.0003***	-0.0005***	0.0002***	-0.109***	0.270***	0.202***
	(0.009)	(0.002)	(0.00005)	(0.00002)	(0.00001)	(0.003)	(0.001)	(0.021)
Income Q2	-0.294***	0.0021	-0.0009***	-0.0004***	0.00001	-0.150***	0.316***	0.267***
	(0.009)	(0.002)	(0.0001)	(0.00001)	(0.00001)	(0.003)	(0.001)	(0.017)
Income Q3	-0.342***	-0.0056**	-0.0024***	-0.0005***	-0.0002***	-0.175***	0.400***	0.402***
	(0.011)	(0.003)	(0.0001)	(0.00002)	(0.00002)	(0.004)	(0.002)	(0.015)
Income Q4	-0.433***	-0.021***	-0.005***	-0.0005***	-0.0007***	-0.223***	0.542***	0.630***
	(0.017)	(0.004)	(0.0002)	(0.00002)	(0.00004)	(0.006)	(0.005)	(0.013)

Notes: This table presents GMM estimates of the preference parameters $\tilde{\beta}_c$ from Equation (7), separately for the four quartiles of income (residual of household size and age and year indicators). This table parallels the estimates in Table 4, except adding convenience and shelf life as additional product characteristics. Shelf life is measured in months per 1,000 calories and top-coded at one year. Convenience is a score per 1,000 calories ranging from 0 to 3, defined as follows. 0: basic ingredients. These are raw or minimally processed foods used in producing a meal or snack that are generally composed of a single ingredient, such as milk, dried beans, rice, grains, butter, cream, fresh meat, poultry, and seafood. 1: complex ingredients, such as bread, pasta, sour cream, sauce, canned vegetables, canned beans, pickles, cereal, frozen meat/poultry/seafood, canned meat/poultry/seafood, and lunch meat. 2: ready-to-cook meals and stacks. These are foods that require minimal preparation involving heating, cooking, or adding hot water, such as frozen entrees, frozen pizzas, dry meal mixes, pudding mixes, soup, chili, and powdered drinks. 3: ready-to-eat meals and snacks. These are foods that are intended to be consumed as is and require no preparation beyond opening a container, including refrigerated entrees and sides, canned and fresh fruit, yogurt, candy, snacks, liquid drinks, and flavored milk. Shelf life data are from Okrent and Kumcu (2016), while convenience data are from the U.S. government's FoodKeeper app (HHS 2015). Magnitudes represent willingness to pay for a unit of the nutrient, where the units are those used in the Health Index. Sodium is in grams; whole fruit, other fruit and dairy are in cups; whole grains, refined grains, and both types of protein are in ounces, added sugar is in teaspoons; solid fats are in calories. "WTP for Health Index" in column 16 equals $\sum_c \hat{\tilde{\beta}}_c s_c r_c$, where s_c is the maximum possible score on the Healthy Eating Index for dietary component c, and r_c is the difference in consumption of component c to receive the maximum instead of the minimum score. Standard errors, clustered by household, are in parentheses. *, **, ***: statistically significant with 10, 5, and 1 percent confidence, respectively.

Table A11: **Model Estimates: Robustness**

Panel (a): Willingness To Pay For Health Index

	(1)	(2)	(3)	(4)
Income Q1	0.43***	0.45***	0.4185***	0.431***
	(0.011)	(0.018)	(0.021)	(.021)
Income Q2	0.63***	0.561***	0.532***	0.557***
	(0.006)	(0.021)	(0.024)	(0.024)
Income Q3	0.82***	0.784***	0.751***	0.762***
	(0.003)	(0.012)	(0.014)	(0.014)
Income Q4	1.14***	1.217***	1.198***	1.194***
	(0.003)	(0.016)	(0.015)	(0.015)
Additional fixed effects:	Baseline	Region × group	Region × rural × group	Region × group × county income

Panel (b): Decomposing the Health Index Gap Across Income Groups

	(1)	(2)	(3)	(4)
Prices (supply)	2.20%	7.80%	1.80%	2.50%
Nutrients (supply)	6.80%	3.90%	5.80%	7.40%
Total supply effect:	9.00%	11.70%	7.60%	9.90%
Nutrient preferences (demand)	37.00%	44.10%	20.00%	27.50%
Product group preferences (demand)	54.00%	44.30%	72.50%	62.60%
Total demand effect:	91.00%	88.40%	92.50%	90.10%
Additional fixed effects:	Baseline	Region × group	Region × rural × group	Region × group × county income

Notes: This table illustrates the robustness of the preference parameters $\tilde{\beta}_c$ from Equation (7) when additional fixed effect controls are added to the estimation. WTP for Health Index" in Panel (a) equals $\sum_c \hat{\tilde{\beta}}_c s_c r_c$, where s_c is the maximum possible score on the Healthy Eating Index for dietary component c, and r_c is the difference in consumption of component c to receive the maximum instead of the minimum score. The decomposition in Panel (b) parallels Figure 6. The baseline specifications in column 1 are as reported in Table 4 and Figure 6, including product group, county, and year fixed effects. Columns 2-4 include additional fixed effects, where "region" is census region, "rural" is an indicator for urban vs. rural county, and "county income" is an indicator for whether the county median income is above the nationwide median county median income. Standard errors, clustered by household, are in parentheses. *, **, ***: statistically significant with 10, 5, and 1 percent confidence, respectively.

Table A12: **Decomposition of Nutrition-Income Relationship by Household Demographics**

	(1)	(2)	(3)
	Full model	Unconditional relationship	Auxiliary regressions
ln(Income)	0.134	0.267	
	(0.0209)***	(0.0178)***	
ln(Years education)	0.685		1.922
	(0.0939)***		(0.0688)***
1(White)	-0.247		0.0588
	(0.0474)***		(0.0283)**
1(Black)	-0.291		-0.155
	(0.0574)***		(0.0342)***
1(Married)	0.0435		0.431
	(0.0256)*		(0.0209)***
Employed	0.0701		0.635
	(0.0802)		(0.0234)***
Weekly work hours	-0.0000934		0.0193
	(0.00225)		(0.000611)***
Health importance	0.102		0.0691
	(0.0119)***		(0.0109)***
Nutrition knowledge	0.132		0.155
	(0.0128)***		(0.0118)***
Household size	-0.103	-0.119	
	(0.0116)***	(0.0111)***	
Census division indicators	Yes	No	No
Age indicators	Yes	Yes	Yes
Year indicators	Yes	Yes	Yes
Observations	81,839	81,839	81,839

Notes: These regressions use 2004–2016 Nielsen Homescan data at the household-by-year level, using only the subsample that responded to the Homescan add-on survey carried out by Nielsen for Allcott, Lockwood and Taubinsky (2018). Health importance is the response to the question, "In general, how important is it to you to stay healthy, for example by maintaining a healthy weight, avoiding diabetes and heart disease, etc.?" Nutrition knowledge is from a battery of 28 questions drawn from the General Nutrition Knowledge Questionnaire (Kliemann et al., 2016). Health importance and nutrition knowledge are both normalized to have a mean of zero and a standard deviation of one. Columns 1 and 2 present estimates of Equation (16), a regression of the Health Index of demand-only consumption predictions on covariates. Each row of column 3 presents the coefficient from a regression of natural log of household income on the variable listed in each row, controlling for age and year indicators and household size. Observations are weighted using the Homescan sample weights. Robust standard errors, clustered by household, are in parentheses. *, **, ***: statistically significant with 10, 5, and 1 percent confidence, respectively.