The Geography of Poverty and Nutrition: Food Deserts and Food Choices Across the United States

Hunt Allcott, Rebecca Diamond and Jean-Pierre Dubé

September 14, 2017

Abstract

We study and quantify the variation in healthful consumption habits across income brackets in the United States. Using a detailed grocery shopping database, we document a positive nutrition-income gradient: the highest-income quartile of households buy groceries that are 0.25 standard deviations more healthful than the lowest-income quartile. We then study the underlying mechanisms leading to the gap. On the supply side, stores in low-income neighborhoods offer fewer produce items and overall less-healthful groceries. However, we find no evidence to support the hypothesis that the low availability of healthful groceries causes unhealthful eating. Using an event study framework, we find that nearby supermarket entry does not increase households’ produce purchases or the overall healthfulness of grocery purchases. Instead, we observe substitution in household expenditures away from incumbent supermarkets. On the demand side, We estimate a formal demand model and use the estimated preferences to decompose differences in purchasing patterns between low- and high-income households. We find that 91% of the nutrition-income gradient is driven by differences in demand across products, while only 9% can be attributed to differences in supply. The demand impact can be broken down as follows: 16% due to product group demand differences, and 75% due to nutrient preferences. The 9% supply effect is entirely driven by different product offerings, with prices actually mitigating the nutrition income relationship by 2%.
I Introduction

Studies by Chetty et al. (2014), Saez and Piketty (2003), and many others have drawn increased attention to the causes and consequences of socioeconomic inequality. These consequences play out in many different ways, including educational opportunities, health outcomes, and involvement with the criminal justice system. In this paper, we study one important correlate of socioeconomic status – what we eat and drink, which in turn affects our health – and quantify the economic mechanisms that drive the nutrition-income relationship.

Obesity is now one of the most important health problems in the U.S. and many other countries. Thirty-five percent of American adults are obese, up from 15 percent in the late 1970s, and an additional 34 percent are overweight (NCHS 2013, 2014).\(^1\) Obesity is estimated to be responsible for 10-27 percent of US medical costs, amounting to several hundred billion dollars annually (Finkelstein et al., 2009; Cawley et al., 2015). Obesity rates are highest among the poor and those with the least education (Drewnowski and Specter, 2004). Low-income women are 45 percent more likely than high-income women to be obese, and women who have not completed college are about 70 percent more likely than those who have (Ogden et al., 2010). It is also well-established that there are socioeconomic differences in what people eat and drink and that these choices are a leading contributor to obesity.\(^2\)

Economists might propose four broad categories of potential explanations for why income could be associated with more healthful eating: availability, prices, preferences, and information. On the supply side, the so-called food desert hypothesis\(^3\) argues that low availability of healthful foods in low-income neighborhoods causes the poor to eat less healthfully. Healthful foods could exhibit lower availability in low-income areas either because of preference externalities or because of higher costs.\(^4\) Although less explored in the extant literature, hypotheses related to the relative price of healthful foods, which could either cost more per calorie or could cost relatively more in low-income areas, argue that the poor choose less healthful food to save money (Drewnowski, 2009).\(^5\) On the demand side, household income could be associated with preferences, either for the taste of healthful foods or for correlates of healthfulness such as perishability. Alternatively, higher-income households could be better informed about the healthfulness of specific foods and the benefits of

---

\(^1\) Overweight and obese, are defined as having Body Mass Index (BMI) of at least 25 and 30 \(\text{kg/m}^2\), respectively, where BMI is weight divided by height squared.

\(^2\) Results from the National Health and Nutrition Examination Survey (NHANES) show that high-income adults get a larger share of their calories from protein and fiber and a smaller share from sugar and cholesterol (USDA 2014a).

\(^3\) Wrigley (2002) attributes the term “food desert” to the Low Income Project Team in the UK, which argued for improved access to retail services in poor neighborhoods.

\(^4\) Recent empirical evidence provides some indirect support for mechanisms that could generate differences in availability: Ellickson (2006, 2007) highlights the role of fixed costs in grocery retail, which are necessary for preference externalities, and people who moved to higher-income neighborhoods in the Moving to Opportunity randomized experiment were less likely to be obese (Kling et al., 2007; Ludwig et al., 2011).

\(^5\) Cutler et al. (2003), Lakdawalla and Philipson (2002), and Philipson and Posner (1999) highlight the importance of prices and argue that cost decreases for unhealthful foods are largely responsible for the obesity increase of the past 40 years.
healthful eating.

Teasing apart these supply and demand side explanations for the nutrition-income relationship is crucial for understanding whether and how policymakers should intervene. The public health literature and public policy debate have given substantial attention to the food desert hypothesis, in spite of limited empirical evidence.\footnote{In an influential review article, Bitler and Haider (2011) write that “it appears that much of the existing research implicitly assumes that supply-side factors cause any food deserts that exist.” The few studies using before/after analyses of retail interventions on diet have found mixed results indicating, at best, a modest effect (e.g., Wrigley et al., 2003; Cummins et al., 2005). Similarly, while the price premium for healthful foods is well-documented (e.g., Drewnowski et al., 2004; Monsivais and Drewnowski, 2007), the evidence for a relatively higher price premium in poor areas is mixed (Acheson, 1998; Kaufman et al., 1997).} The US Healthy Food Financing Initiative allocates $125 million annually in grants, loan subsidies, and technical assistance for grocers in underserved areas.\footnote{In the U.S., the Healthy Food Financing Initiative has awarded over $140 million since 2011 through a suite of programs that finance and provide technical assistance to grocery stores, farmers markets, and other suppliers of healthful foods in under-served areas (TRF 2015a). Pennsylvania’s Fresh Food Financing Initiative provided $85 million grants and loans to retailers offering fresh foods in under-served low-income areas (TRF 2015b). Projects aimed at “eliminating food deserts” were eligible for the $100 million in Community Transformation Grants under the Affordable Care Act. Similarly in the United Kingdom, the 2001 Food Poverty Eradication Bill required local and national governments to document and take actions to eliminate food deserts.} Such policies could increase both efficiency and equity if they address market failures, reduce costs, and affect purchases. But if the nutrition-income relationship is primarily driven by preference differences, it is less clear that improving availability will have any effect, let alone improve welfare.

We combine reduced-form analyses with a structural demand model to quantify the relative importance of prices, availability, preferences, and information in generating the nutrition-income relationship. We exploit a rich combination of datasets including Nielsen Homescan (HMS), a 60,000-household, nationally-representative panel survey of grocery consumption, and Nielsen’s Retail Measurement Services (RMS), a 35,000-store, national panel of product-level sales data. Therefore, unlike past survey-based research, we observe households’ food quantity choices, expenditures and choice sets. We match these marketing data with each product’s nutritional content, perishability, and convenience of preparation. We also match the household data with Nielsen Panelviews survey-based measures of household-level nutritional knowledge and information sets. Finally, we include geocoded data on the entry of new supermarkets.

We begin by documenting several relevant stylized facts. First, there is a meaningful nutrition-income relationship in grocery purchases: the top income quartile buys groceries that are 0.29 standard deviations more healthful than the bottom quartile, as measured by our Health Index, which scores purchases based on U.S. government recommendations. One of our key objectives consists of studying the sources of this difference. During the sample period, the gap between high and low income households increases substantially, highlighting the relevance of our analysis. As in Handbury et al. (2015), we find that education explains about one third of the nutrition-income relationship. However, less than half of the nutrition-income correlation is explained by standard household demographics, suggesting unobserved aspects of preferences or supply-side factors may be important. Second, healthful food costs more for all households. This effect is
driven by the cost of produce. However, we do not find a higher price premium for healthful foods in low-income areas. On the contrary, we find produce is actually 2% cheaper in low-income areas. We find no price premium for other non-produce healthful foods. (as in Kaufman et al., 1997). Third, the average store in low-income neighborhoods offers less healthful groceries than in high-income neighborhoods. This difference is almost entirely explained by the fact that there are fewer “supermarkets” (by which we mean large grocery stores, supercenters, and club stores) and more drug and convenience stores in low-income neighborhoods. However, fourth, even low-income households buy most of their groceries from supermarkets. The relatively high geographic density of supermarkets ensures that the mean (median) shopping trip is five (three) miles one-way, which moderates the role that local neighborhood supply conditions play in constraining choice sets. In sum, we observe several factors on both the supply and demand sides that could potentially contribute to the nutrition-income gap. However, the willingness-to-drive works against the food desert hypothesis, suggesting that increased retail access may not have much effect on healthful food choices in poor neighborhoods.

To test the food desert hypothesis and the role of availability on healthful choices, we exploit the impact of new supermarket entry on within-household purchase behavior. Unlike past work, we observe a household’s shopping basket for all trips across all store formats. Consistent with past work (e.g., Wrigley et al., 2003; Cummins et al., 2005), we find no effect of new store entry on the healthfulness of a household’s purchases. We find tightly-estimated zero effects on two measures of healthful purchases: the share of calories from produce and the Health Index of all grocery purchases. In addition, we are able to document a substantial substitution effect in grocery expenditures towards the new local supermarket chain entrant. However, these expenditures are primarily substituted from other supermarkets, not from drug stores and convenience stores – even for households living in food deserts (i.e., zip codes with no supermarkets). In sum, store entry likely reduces transport costs, but does not have a substantial effect on the composition of the shopping basket itself. We can bound the effects of supermarket entry within a 10-minute drive at no more than 1/6 the inter-income-quartile range of produce calorie share and no more than 1/32 the inter-income-quartile range of Health Index. Thus, having fewer supermarkets nearby explains essentially none of the nutrition-income relationship.

To test for and quantify the role of heterogeneous nutrition preferences on the demand side, we derive a demand model based on Dubois et al. (2014). The model allows for Constant Elasticity of Substitution (CES) preferences over individual products defined at the Universal Product Code (UPC) level, combined with Cobb-Douglas preferences for product groups (milk, breads, candy, vegetables, etc.) and aggregate preferences over specific macronutrients (saturated fat, sugar, salt, etc.).

To estimate the model, we develop a new instrument for prices driven by variation in different grocery chains’ comparative advantages in pricing different products. Using geographic in presence of each chain shifts which product categories are particularly cheap in each market. To illustrate,
consider a simple example in which there are two types of foods – apples and pizza – and two
grocery chains – Safeway and Shaws. Suppose Safeway is able to source pizza cheaply, while Shaws
can source apples cheaply. Then, cities dominated by Safeway will face relatively low prices for
pizza and cities dominated by Shaws will face relatively low prices for apples. This instrument
demostrates a very strong first stage (F-stat > 85), even when including product category and
market-level fixed effects. This instrument may be useful to other researchers using Nielsen data
who would like exogenous price variation.

The demand estimates show that preferences for healthful macro nutrients strongly increase as
one move’s up the income distribution. For example, low income households are willing to pay $5.96
to consumer a kilogram of protein instead of a kg of carbohydrates, high income households are
willing to pay $7.32, 23% more. However, high income households still enjoy unhealthful nutrients.
They are willing to pay $12.50 for a kg of saturated fat versus a kg of carbohydrates. Nonetheless,
low income households are willing to pay even more ($15.91) for a kg of saturated fat versus
cholesterol. We find similar patterns for all other healthful nutrients (fiber, fruit, vegetables) and
most unhealthful nutrients (sugar, sodium). We find no heterogeneity in willingness to pay for
cholesterol.

Finally, we use our demand model to decompose the nutrition-income gradient into causes
due to supply versus household preferences. We look at several counter-factual scenarios in which
households from the bottom income quartile are exposed to the product availability and prices for
households in the top income quartile. We find that 91% of the nutrition-income gradient is driven
by differences in demand across products, while only 9% can be attributed to differences in supply.
The demand impact can be broken down into 16% due to product category demand differences,
and 75% due to nutrient preferences. The 9% supply effect is entirely driven by different product
offerings, with prices playing essentially no role.

Our paper contributes to a broad literature on the economics of nutrition and health (see Cawley,
2015, for a review). Anderson and Matsa (2011), Currie et al. (2010), Davis and Carpenter (2009),
and Dunn (2010) study the effects of proximity to fast food restaurants on food consumption and
obesity, complementing our analysis of supermarket entry. Their results are qualitatively consistent
with ours in that they suggest that the causal impact of unhealthful food supply is small, relative
to either the overall obesity rate or the nutrition-income relationship. Courtemanche and Carden
(2011) find that the entry of Walmart Supercenters increases BMI and obesity. Our results that
supermarket entry does not significantly affect nutritional content per calorie consumed supports
their interpretation that the mechanism is reduced prices per calorie. However, Volpe et al. (2013)
find that supercenter expansion reduces the healthfulness of grocery purchases, a result that differs
from ours for a number of potential reasons. Relative to their paper, we benefit from having precise
dates and geocoded locations of store entry for several different supermarket chains, which allows
more precise identification.

Our reduced form analysis in Section IV is broadly comparable to the work of Handbury et al.
(2015), and our results echo their conclusion that grocery store entry does not significantly affect healthful eating. We add to this finding by showing why entry has little effect on healthfulness – households travel long distances to shop, so supermarket entry largely diverts expenditures from other supermarkets. Unlike Handbury et al. (2015), who “show that spatial disparities in access play a limited role in generating socioeconomic disparities in nutritional consumption” (page 1), we seek to understand and decompose the underlying potential causes of the nutrition-income relationship. We use a structural approach that uses demand estimates to decompose the sources of the nutrition-income relationship.

In the public health literature, Larson et al. (2009) review 54 studies documenting differences in food access across neighborhoods. These studies are typically either detailed inventories of product availability at specific retailers in specific local areas (e.g., Sharkey et al., 2010) or nationwide studies of store counts by neighborhood income without detailed product data (e.g., Powell et al., 2007). Even in establishing stylized facts, the Nielsen HMS and RMS data in our paper and in Handbury et al. (2015) are groundbreaking in that they provide both large nationwide samples and detailed product availability and purchase information. A few studies in the public health literature do examine food consumption before vs. after supermarket entry (e.g., Wrigley et al., 2003; Cummins et al., 2005, 2015; Elbel et al., 2015). However, according to a major report by the U.S. Department of Agriculture (Ploeg et al., 2009, page v), “the findings are mixed,” perhaps because the standard in this literature has been to study entry of one retail establishment, which limits statistical power and generalizability and makes it more difficult to establish a credible counterfactual. By contrast, our paper and Handbury et al. (2015) evaluate the effects of thousands of supermarkets as they enter and exit across the U.S. during a ten-year period.

The remainder of the paper is organized as follows. Sections II through VIII, respectively, present data, stylized facts, reduced form empirical analysis, structural demand model, structural estimates, and the conclusion.

II Data

II.A Nielsen Homescan and Retail Scanner Data

We use the Nielsen Homescan Panel (HMS) for 2004-2015 to measure household grocery purchases.\(^8\) HMS includes about 39,000 households each year for 2004-2006, and about 61,000 households each year for 2007-2015. HMS households record UPCs of all consumer packaged goods they purchase from any outlet. We consider only food and drink purchases, and we further exclude alcohol and health and beauty products such as vitamins.

We focus on explaining income-related differences in grocery purchases, not overall diets, because HMS does not include data on food purchased in restaurants.\(^9\) One additional limitation of HMS

---

\(^8\) See Einav et al. (2010) for a validation study of Homescan.

\(^9\) The National Health and Nutrition Examination Survey (NHANES) finds that Americans consume 34 percent of calories away from home, including 25 percent in restaurants (USDA 2014b). For all income groups, the share of
is that most households only record purchases of packaged items with UPCs, not non-packaged groceries such as bulk produce and grains. For 2004-2006, however, the data also include an 8,000-household “magnet” subsample that also recorded prices and weights of non-packaged groceries. We use the magnet data for robustness checks.\footnote{The magnet data continue after 2006, but panelists now record only expenditures and not weights purchased. Because prices per unit weight can vary substantially across stores and neighborhoods, we do not use these data to construct food purchases.} Appendix Figure A1 shows that about 60 percent of magnet households’ produce calories are from packaged goods that are observed in the full HMS sample, and this proportion does not vary statistically by income.\footnote{After excluding canned, frozen, and dried produce, about 39 percent of magnet households’ fresh produce calories are from packaged items. Households with incomes less than about $20,000 buy about five percentage points less of their fresh produce calories from packaged items, but the proportion is constant at moderate and high incomes.}

HMS households report demographic variables such as household income in 16 bins, presence of children, race, and the age, educational attainment, employment status, and weekly work hours for male and female household heads. In the usual case where there are two household heads, we use the mean of age, education, and employment variables observed for both heads. The U.S. government Dietary Guidelines include calorie needs by age and gender; we combine that with HMS household composition to get each household’s daily calorie need. In addition to the standard HMS data, we observe occupation, results of a food knowledge quiz, and state of birth from add-on surveys carried out by Bronnenberg et al. (2012) and Bronnenberg et al. (2015). Panel A of Table 1 presents descriptive statistics for HMS households. Unless otherwise stated, all HMS results are weighted for national representativeness.

The Nielsen Retail Measurement Services (RMS) data consist of weekly prices and sales volumes for each UPC sold at approximately 42,000 stores at 160 retail chains for 2006-2015. We exclude liquor stores. RMS includes 53, 32, 55, and 2 percent of sales in the grocery, mass merchandiser, drug, and convenience store channels, respectively. As with HMS, RMS does not include sales of bulk produce and other non-packaged items.

We gather zip code median income from the American Community Survey 2011 five-year estimates and county mean income from the Regional Economic Information System. We deflate prices to 2010 dollars using the consumer price index for urban consumers for all items.

II.B Grocery Retail Establishments

Studying the effects of retailer entry requires reliable data on store open dates to avoid attenuation bias. Some datasets, such as InfoUSA and the National Establishment Time Series, might be reasonable for cross-sectional analyses, but they do not sufficiently precisely record the open dates of new establishments; see Bitler and Haider (2011, page 162) for further discussion. Furthermore, to measure true changes in availability experienced by consumers, we must use actual new establishments, not store locations that continue to operate but change management.
In light of these issues, we measure entry with two datasets. First, for the period between January 2004 and December 2013, we gathered the exact store open dates and addresses for 1,914 large grocery stores and supercenters spanning multiple chains. Second, we use Zip Code Business Patterns (ZBP), which gives a count of establishments by NAICS code and employment size category for every zip code as of March 10th of each year. The ZBP data are drawn from tax records, the U.S. Census Company Organization Survey, and other administrative data. Panel B of Table 1 presents ZBP descriptive statistics.

II.C Nutrition Facts and the Health Index

We purchased UPC-level nutrition facts from Gladson, and we gathered nutrition facts for non-packaged items from the USDA National Nutrient Database for Standard Reference (USDA 2014c). Panel C of Table 1 presents nutritional summary statistics across all UPC codes in the Nielsen HMS and RMS data.

In parts of the paper, it will be useful to characterize goods and preferences using a one-dimensional index of healthfulness. The most natural option is the USDA’s Healthy Eating Index (HEI), which was designed to score entire diets on an easily-understandable range from 0-100. However, the HEI has sharp non-linearities because it values a balanced diet, so an item’s contribution to the HEI depends on the consumer’s full diet. This is inappropriate for datasets like HMS and RMS where consumers’ full diets are not observed.

We therefore construct a modified version of the HEI that is based on the same U.S. government dietary recommendations but is linear and additively separable in macronutrients. The U.S. government Dietary Guidelines are clearly organized around “good” macronutrients to “increase” vs. “bad” macronutrients to “reduce.” According to the Dietary Guidelines, the “good” macronutrients are fruits, vegetables, protein, and fiber, while the “bad” macronutrients are saturated fat, sugar, sodium, and cholesterol. Our raw Health Index \( H(x) \) is the sum of good minus bad nutrients in a UPC, weighting each by its recommended daily intake (RDI):

\[
H(x) = \sum_k G_k \frac{g_k}{r_k} - (1 - G_k) \frac{g_k}{r_k},
\]

where \( g_k \) is the grams of macronutrient \( k \), \( r_k \) is the RDI for a normal adult, and \( G_k \) takes value 1 for “good” macronutrients and 0 for “bad” macronutrients. For example, the RDI for vegetables is about 130 grams per day, so a vegetable UPC weighing 130 grams would have \( H(x) = 1 \). Packaged items in the Nielsen “candy” product group have an average \( H(x) \) of -5.0. For the model-free analyses in Sections III and IV, we normalize the raw Health Index per calorie consumed to have mean zero and standard deviation one across households after conditioning on year indicators.

Collapsing to one all-encompassing measure of healthy eating risks obscuring important results on other more specific measures. We will also present results using other measures, and when we do focus on Health Index, Appendix Table A2 shows that the strongest correlates of Health Index in household-by-year HMS data are sugar (with correlation coefficient \( \rho \approx -0.75 \)), fiber (\( \rho \approx 0.64 \)), protein (\( \rho \approx 0.52 \)), and fruits and vegetables (\( \rho \approx 0.34 \) and 0.34). See Appendix A.B for additional details on the Health Index.
III  Stylized Facts: Purchases and Supply of Healthful Foods

III.A  The nutrition-income relationship

At the national level, we observe health consumption disparities across US neighborhoods. Figure 1 presents a map of the estimated Health Index of packaged grocery purchases by county, using the 2006-2015 RMS data.\textsuperscript{12} We observe substantial variation in healthfulness across counties that appears to be highly correlated with county mean income (correlation coefficient $\rho \approx 0.53$). The health Index of grocery purchases is also highly correlated with Chetty et al. (2016)’s county-level life expectancy measure ($\rho \approx 0.53$) as well as such health-related behaviors as smoking ($\rho \approx -0.41$), obesity rates ($\rho \approx -0.47$) and exercise ($\rho \approx 0.56$).

At the more micro level of the household, the basic result that eating patterns vary by income has been well-documented through the NHANES survey United States Department of Agriculture (2014), which began in the 1960s. Using the 2004-2015 HMS panel data, Figure 2 presents four different measures of the nutrition-income relationship: share of total calories from sugary drinks (such as soft drinks, pre-made iced tea, and energy drinks), share of bread calories from whole grain breads, share of total calories from (packaged) produce, and normalized Health Index across all grocery purchases. These four measures paint a consistent picture: low-income households purchase more sugary drinks, less whole-grain breads, less produce, and lower-Health Index groceries.\textsuperscript{13} Interestingly, the relationship between the Health Index and income has a steeper slope for household incomes below $20,000. This finding is reminiscent of (and potentially one contributor to) the Chetty et al. (2016) finding of a sharper drop in life expectancy for people in the lowest income percentiles.

Households with income above $70,000 (approximately the top 34 percentiles) buy groceries with a Health Index 0.29 standard deviations higher than households with incomes below $25,000 (approximately the bottom 24 percentiles). One of our key objectives consists of explaining this 0.29 standard deviation difference.

\textsuperscript{12}Since the RMS do not contain the complete census of stores, the distribution of store channel 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 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 area $n$ in RMS vs. the true distribution of store channel types observed in Zip Code Business Patterns.

\textsuperscript{13}Appendix Figure A2 presents analogues to Figure 2, considering each individual macronutrient. Consistent with NHANES, the most economically and statistically significant relationships in HMS are that higher-income households tend to get a larger share of calories from protein and fiber and a smaller share from sugar. Higher-income households also tend to consume more saturated fat, sodium, and cholesterol, although the difference is considerably smaller in percent terms. Thus, although the Recommended Daily Intakes impose specific weights on macronutrients in our Health Index, higher-income diets would tend to be classified as “more healthy” unless the weights change substantially.

Appendix Figure A3 re-creates the figure using the magnet subsample for 2004-2006, which includes bulk purchases as well as packaged items. The results are qualitatively similar, although the income group differences are attenuated slightly because the nutrition-income relationship is less stark in 2004-2006 compared to the full 2004-2015 sample, as shown below in Figure 3.
Past work has reported static estimates of the nutrition-income relationship. Our 12-year panel data sample period allows us to examine potential trends. Figure 3 illustrates that the Health Index increased by 0.20 standard deviations for 2011-2015 relative to 2004-2007 for households with incomes above $70,000. By contrast, the Health Index increased by only 0.04 standard deviations for households with incomes below $25,000. This growing trend in the nutrition-income relationship, especially for higher-income households, underscores the increasing relevance of the issues we study.

To benchmark the potential importance of these differences, note that over the full 2004-2015 sample, households with income above $70,000 purchase approximately one additional gram of fiber and 3.5 fewer grams of sugar per 1000 calories relative to households with income below $25,000. Using correlational analysis, Montonen et al. (2003) find that consuming one additional gram of fiber per 1000 calories is conditionally associated with a 9.4 percent decrease in type-2 diabetes, and results from Yang et al. (2014) imply that 3.5 fewer grams of sugar per 1000 calories is conditionally associated with a ten percent decrease in death rates from cardiovascular disease.

We have also estimated the short-term association between Health Index and household income by conditioning on household-by-location fixed effects and other time-varying household demographics; see Appendix Table A3. Under the assumption that within-household income changes are exogenous and free of measurement error, the results would imply that short-term income changes explain 10-15 percent of the correlation between Health Index and household income. This at least suggests that the bulk of the nutrition-income relationship is explained by long-term effects of income and correlates of income, not by the short-term effects of a tighter budget constraint.

III.B Healthful grocery availability by neighborhood and store format

The food desert hypothesis argues that the nutrition-income relationship is driven by low availability of healthful foods in low-income neighborhoods. We examine this supply-side theory by measuring the healthfulness of the choice sets available in different neighborhoods. The left four panels of Figure 4 present the relationship between local median income and the healthfulness of items offered in RMS stores, again measured by the share of all UPCs that are sugary drinks, share of bread UPCs that are whole grain, share of all UPCs that are produce, and mean Health Index of UPCs offered. Because this figure weights UPCs only by calories in the package and not by quantity sold, this figure reflects choice sets, not consumption. All four panels show the same qualitative result: stores in higher-income zip codes offer healthier items. This pattern of less

14 Appendix Figure A4 shows that wealthy households’ differential increases in Health Index are driven by improvements in nearly all macronutrients that comprise the Health Index: relative to low-income households, high-income households have increased purchases of produce, protein, and fiber, and decreased purchases of sugar and sodium, over the 2004-2015 sample period. The one exception is cholesterol, an unhealthy macronutrient that high-income households purchase relatively more of later in the sample.

15 Unlike the nutrition-income relationship in consumption from Figure 2, the weights in the Health Index could matter for the conclusion that choice sets are more healthful in higher-income neighborhoods. Appendix Figure A5 shows that while stores in higher-income neighborhoods offer UPCs with more protein and fiber and less sugar per 1000 calories, they also offer more of three “bad” macronutrients, saturated fat, sodium, and cholesterol.
healthful offerings in low-income neighborhood stores is broadly consistent with the public health literature (e.g., Larson et al., 2009) and with Handbury et al. (2015)’s analysis of the RMS data.

The two panels at the right of Figure A7 show that RMS stores in low-income neighborhoods are also significantly smaller and offer considerably less variety. The mean store in zip codes with median household income below $25,000 offers 4,200 UPCs, while the mean store in zip codes with median household income above $70,000 offers 9,800 UPCs.

Table 2 formalizes these correlations in store-level regressions using the 2006-2015 RMS data. We consider two measures of healthy grocery availability at store \( j \) in year \( t \), \( H_{jt} \): the count of produce UPCs offered and the calorie-weighted mean Health Index of UPCs offered. The table presents regressions of \( H_{jt} \) on the natural log of zip code median income and additional store covariates \( X_{jt} \):

\[
H_{jt} = \alpha \text{Zip Median Income}_j + \beta X_{jt} + \varepsilon_{jt}
\]  

Columns 1 and 4 confirm that that stores in higher-income zip codes offer substantially more produce UPCs and overall healthier items.

We consider two potential explanations for the lower healthfulness of the choice sets in low-income zip codes. First, we look at store size using All Commodity Volume (ACV), the total annual revenues of the store.\(^{16}\) Since ACV is partially driven by prices, we also use the total number of unique UPCs in a year as a proxy for total shelf space. The rightmost panels of Figure 4 indicate that stores in higher-income neighborhoods are considerably larger in terms of ACV and in terms of total UPCs sold than stores in low-income neighborhoods. Larger stores naturally offer more variety and could offer healthier UPCs. Second, even after conditioning on store size, stores in low-income neighborhoods could systematically stock less healthy options.

We test between these two explanations in Columns 2 and 5. ACV explains almost all of the relationship between neighborhood income and healthy grocery availability. In particular, large stores systematically offer healthier groceries and are much more prevalent in higher-income zip codes. Columns 3 and 6 show that, even when excluding the ACV variable, retail “channel type” indicators explain 75 to 80 percent of the income-healthfulness relationship. In separate regressions, we find that channel type indicators alone explain 92 and 73 percent of the variance in produce UPC counts and mean Health Index, respectively.

Table 2 also shows that grocery stores, supercenters, and club stores offer more produce UPCs and a higher average Health Index than convenience stores, drug stores, and other mass merchants. Indeed, supercenters such as Walmart Supercenter, Target, and Meijer by definition have full lines of grocery and produce items, as do club stores such as Sam’s Club and Costco. In the analysis below, we thus refer to large grocers, supercenters, and club stores as “supermarkets,” with the understanding that supermarkets carry more produce and more healthful items. We define a “food

---

\(^{16}\)The retailing literature typically uses ACV to measure store size (Hoch et al., 1995), a practice developed by the retail industry to assess how much business is conducted by the store (Martin, 2015).
Consistent with Figure 4 and Table 2, Appendix B.B shows that low-income neighborhoods indeed tend to have fewer large grocery stores and more drug and convenience stores per capita. The ZBP data show that 24 percent of zip codes (weighted by population) are “food deserts.” By contrast, 55 percent of zip codes with median income below $25,000 are food deserts. This lower concentration of healthy stores has generated support among policymakers for incentivizing entry of new supermarkets into low-income neighborhoods. We will estimate the effects of this entry in Section IV.A.

In sum, the relationship between local income and the healthfulness availability in stores is mainly driven by differences in the store formats present across zip codes. Low-income zip codes tend to have smaller store formats that typically stock less healthful food options. These results have important implications. First, income disparities in healthy grocery availability are driven by the entry decisions of specific firms, not product stocking decisions. Second, store format explains most of a given store’s healthy product offerings, potentially obviating the need for more detailed sales data. By extension, the count of observed store formats from sources like the Zip Code Business Patterns reveals most of the neighborhood-level availability of healthy foods. Collectively, these findings suggest a potential supply-side explanation for the nutrition-income relationship in household purchases.

III.C The healthful food price premium

Public health studies typically find that healthful diets cost more per calorie; see Rao et al. (2013) for a meta-analysis of 27 studies. We take all the RMS grocery sales for 2012 and divide them into three groups: produce, above-median Health Index non-produce, and below-median Health Index non-produce. Figure 5 presents the mean price per calorie for UPCs stocked in each of these three groups, where we weight the RMS stores based on the trips taken to them by each household income category. Broda et al. (2009) show that higher-income HMS consumers pay slightly more for the same UPCs. The upward slope of price per calorie in income reflects that finding, stores frequented by low income households charge 2.15% less for produce, 0.85% less for healthful non-produce and 0.93% less for unhealthful foods than stores frequented by high income households. We define unhealthful as product groups with below medium health indices. This figure quantifies two additional facts. First, across all income groups, produce cost more per calorie than relatively non-produce UPCs. This could cause low-income households to purchase less produce (less healthful food) in order to more cheaply satisfy minimum calorie needs. Second, however, the relative price of produce and healthful non-produce UPCs increases in income. As Handbury et al. (2015) point out, this relative price difference should cause low-income households to purchase more healthful groceries. Our demand model will allow us to quantify the effects of these price differences.
IV Effects of Retailer Entry and Local Food Environments

Having documented the presence of food deserts and a systematically lower availability of healthful foods in low-income neighborhoods, we now test whether this availability on the supply side is driving nutritional choices. Our key identification challenge is that cross-sectional differences in availability could also reflect systematic differences in household types due to geographic sorting. We therefore use a within-household approach that compares choices before and after the entry of a nearby supermarket. To test whether the neighborhood itself may be driving choices, we also use a within-household approach that compares choices before and after a household moves across zip codes and counties within the sample period.

IV.A Effects of Supermarket Entry

We use household-by-quarter HMS data in an event study framework to measure the effects of supermarket entry on grocery purchases. Using the google maps application program interface (API), we downloaded the driving time (assuming no congestion delay) between each Census tract centroid and the address of each of our 1,914 entering supermarkets.\(^\text{17}\) \(S_{mct}\) is the count of supermarket entries that have previously occurred within driving distance band \(m\) of Census tract \(c\) as of quarter \(t\) or earlier. We use two distance bands, \(m \in [0, 10)\) minutes and \(m \in [10, 15)\) minutes. 10 and 15 minutes are the median and 75th percentile of shopping travel times in the 2009 NHTS. Almost all households experience either zero or one entry in our data: for \(m \in [0, 10)\) minutes, \(S_{mct}\) takes value 0 for 87 percent of observations, 1 for 11 percent of observations, 2 for 1.4 percent of observations, and 3, 4, or 5 for the remaining 0.3 percent. Since the set of households exposed to local entry are not nationally representative, we do not use the HMS sample weights for this analysis.

Let \(X_{it}\) denote the vector of potentially time-varying household covariates presented in Table 1: natural log of current-year income, natural log of years of education, indicators for each integer age from 23-90, an indicator for the presence of children, race indicators, a married indicator, employment status, weekly work hours, and household daily calorie need. Let \(Y_{it}\) denote an outcome for household \(i\) in quarter \(t\).

We then run the following regression:

\[
Y_{it} = \sum_m \tau_m S_{mct} + \beta X_{it} + \mu_{dt} + \phi_{ic} + \epsilon_{it}
\]

where \(\mu_{dt}\) is a vector of Census division-by-quarter of sample indicators, and \(\phi_{ic}\) is a household-by-Census tract fixed effect. As we study in Section IV.B, some HMS households move while in the sample. So \(\phi_{ic}\) isolates variation in supply due to entry, not relocation. The coefficients,\(^\text{17}\)

\(^{17}\)In our 2015 working paper, we used geographic distances, as do Handbury et al. (2015). Because driving five miles in a dense city takes much more time than driving five miles in a suburban area, drive time is a more precise measure of travel costs.
\( \tau_m \), measure the effect of entry under the identifying assumption that store entry is exogenous to within-household preference changes over time. While retailers carefully plan entry and exit in response to local population growth and changes in local demographics, we assume that entry is uncorrelated with within-household demand changes conditional on division-by-quarter fixed effects and household demographics.

Before estimating Equation (2), we first show graphical results of the event study. We define \( E_{cmqt} \) as an indicator variable for whether one supermarket entered within distance \( m \) of Census tract \( c \), \( q \) quarters before quarter \( t \). \( B_{it} \) is an indicator variable for whether observation \( it \) is part of a balanced panel around one supermarket entry: the household is observed in the same Census tract continuously for all three quarters before and all eight quarters after one supermarket entry, with no other entries during that period. We run the following regression:

\[
Y_{ict} = \sum_{m} \tau_{mq} E_{cmqt} B_{it} + \alpha B_{it} + \beta X_{it} + \mu_{it} + \phi_{ic} + \varepsilon_{ict}.
\] (3)

Figure 6 presents the \( \tau_{0,10q} \) coefficients and 95 percent confidence intervals, illustrating how one supermarket entry within ten minutes of household \( i \) affects purchases in event time, using the balanced panel around the entry. The omitted category is \( q = -1 \), so all coefficients are relative to the outcome in the last quarter before entry. Standard errors are clustered by household.

The top two panels present expenditure shares (in units of percentage points), showing that entry clearly affects purchasing patterns. The dependent variable in the top left panel is the combined expenditure share across the several retail chains we observe in our entry dataset. Expenditures at entering retailers increase by about six percentage points, and most of the adjustment occurs within the first four quarters after entry.\(^\text{18}\)

This approach is likely to understate the expenditure share increase at the specific entering store which will divert expenditures from other stores owned by the same retailer in addition to the other retailers in our entry dataset. The top right figure shows combined expenditures at grocery stores, supercenters, and club stores, which as we have shown above tend to offer a wider variety of produce and healthier items overall. To the extent that supermarket entry simply diverts sales from these store types, the actual changes in product availability (and thus the possible effects on healthful purchases) are more limited. Indeed, the combined expenditure share at supermarkets increases by only one percentage point by the eighth quarter after entry. Thus, the primary effect of supermarket entry is to divert sales from other supermarkets.\(^\text{19}\)

The bottom two panels present results with Health Index of purchased groceries as the dependent variable. The bottom left panel uses the full sample, while the bottom right panel restricts the sample to the 25 percent of the sample that lives in “food deserts”—that is, zip codes with

\(^\text{18}\)This gradual adjustment of purchases is consistent with the results of Atkin et al. (2016), who study retail expansion in Mexico.

\(^\text{19}\)These results are consistent with those of Hwang and Park (2015), who look at a subset of Walmart Supercenters that opened between 2003 and 2006.
no supercenters, club stores, or large (>50 employee) grocery stores as of 2003. Both figures show no increase in healthy food purchases after supermarket entry. Appendix Figure A8 presents the analogous figures for the $\tau_{10,15,q}$ coefficients. The expenditure share changes are attenuated, as would be expected given that the entering stores are 10-15 minutes away instead of 0-10 minutes away, and there is again no apparent change in Health Index.

Table 3 presents estimates of Equation (2), using the same dependent variables as Figure 6. Standard errors are clustered by household. Panel A considers the effects on expenditure shares, first at the entrant retailers and then at all grocery stores, supercenters, and club stores. Unsurprisingly, all effects are significantly larger for stores entering within ten minutes of a household’s Census tract centroid than for stores entering 10-15 minutes away. Columns 1 and 2 consider the full HMS sample, columns 3 and 4 limit the sample to low-income households, and columns 5 and 6 limit to households in “food deserts.” The expenditure share changes are generally larger for low-income households and households in food deserts. However, consistent with Figure 6, 75 to 85 percent of entrant chains’ expenditure share increase consists of diverted sales from other grocery stores, supercenters, and club stores, while less than one-quarter is diverted from the other store channel types that typically offer less healthy groceries. Appendix Table A4 shows that most of this diversion from less healthy channel types is from other mass merchants; even in food deserts, supermarket entry causes expenditure shares at drug and convenience stores to drop by only a fraction of a percentage point.

Panel B of Table 3 presents effects on Health Index, which again is normalized to standard deviation 1 across households. Consistent with Figure 6, entry has no statistically significant effect, with one exception: there is a statistically significant but economically small effect of supermarket entry 10-15 minutes away from households in food deserts. Appendix Table A4 repeats these estimates using three alternative definitions of food deserts; again, there is a tightly estimated zero effect of entry within 0-10 minutes and a statistically significant but economically small effect of entry 10-15 minutes away.

We can use these estimates to determine the share of the difference in Health Index between low-income and high-income households that can be explained by having more local supermarkets. From Section III.A, we know that households with income above $70,000 buy groceries with Health Index 0.29 standard deviations higher than those with income below $25,000. The upper bound of the 95 percent confidence interval from column 2 of Panel B implies that one supermarket entry increases Health Index by no more than 0.036 standard deviations for low-income households. Using the Zip Code Business Patterns, we calculate that the same high-income HMS households have an average of 2.4 supermarkets in their zip code, while the low-income households have an average of 2.0, for an average difference of 0.4 supermarkets. Thus, we can conclude that local access to supermarkets explains no more than $0.036 \times 0.4/0.29 \approx 4.8\%$ of the high- vs. low-income difference in Health Index.

These regressions consider entry by only a limited set of retailers, which could reduce both
power and external validity. Appendix C.B presents parallel estimates using zip code-by-year counts of large grocery stores, supercenters, and club stores from Zip Code Business Patterns. These estimates are closely consistent with the results in Table 3: even in food deserts, the main effect of supermarket entry is to divert expenditures from other supermarkets, so consumers’ choice sets are largely unchanged. Entry has tightly-estimated zero effects on Health Index for the full sample as well as the low-income and “food desert” subsamples. In sum, differences in availability of healthful products – the so-called food desert hypothesis – do not appear to be driving the nutrition-income relationship in household purchases.

IV.A.1 Why Doesn’t Entry Matter?

Especially given the academic and policy attention to local access to healthy grocery options, it is remarkable that supermarket entry seems to matter so little. We document two key facts to help explain the lack of an entry effect.

First, Americans travel long distances to shop, typically in cars. This fact can be seen with the 2009 National Household Travel Survey (NHTS), a nationally-representative survey that gathers demographics, vehicle ownership, and “trip diaries” from 150,000 households. Figure 7 shows average one-way distances for trips beginning or ending in “buying goods: groceries/clothing/hardware store.” The mean trip is 5.2 miles, the median trip is 3.0 miles, and 90 percent of shopping trips are by car. People with household income less than $25,000 (labeled as “poor” on the figure) travel a mean of 4.8 miles. People who live in “food deserts,” again defined as zip codes with no large grocery stores, supercenters, or club stores at the time of the NHTS survey, travel almost seven miles on average.

The HMS sample is limited to urban areas, where travel distances may be shorter. The right of Figure 7 thus excludes rural areas (Census places with population less than 2500), showing that mean travel distances are still 4.5 miles, and 5.0 miles in urban food deserts. Low-income households who live in food deserts and do not own a car, a disadvantaged group that would be heavily affected by lack of local access to healthy groceries, represent only 0.6 percent of the population and travel a mean of 2.0 miles. Appendix Figure A6 presents median travel distances and the share of trips by auto for the same subgroups. These long travel distances suggest that people may still shop in supermarkets even if they don’t have a supermarket nearby. However, these travel distances are only suggestive, as the NHTS trip diaries do not gather information on store channel types.

Second, Figure A9 shows that low-income households and households in food deserts spend only slightly less in supermarkets. Similarly, Broda, Leibtag, and Weinstein (2009) report relatively small expenditure share differences across income groups. Households with income below $25,000 spend about 87 percent of their grocery dollars at supermarkets, while households with incomes above $70,000 spend 91 percent. For households in our “food deserts,” the supermarket expenditure share is only a fraction of a percentage point lower.20

20Appendix Figure A9 presents expenditure shares by income category for all channel types.
Clearly, households in food deserts may benefit from the increased store variety and reduced travel costs associated with supermarket entry. However, because most consumers already travel to shop in supermarkets, supermarket entry does not significantly change choice sets, and thus doesn’t affect healthy eating.

IV.B “Place Effects” Identified by Movers

A potential concern with our lack of a supermarket entry effect on healthful food choices a food desert is that some broader place effects drive purchase decisions. For instance, peer effects from the healthful eating habits of friends and neighbors as well as general local knowledge and image concerns related to healthy eating could drive a household’s choices.

To test for place effects, we look at the within-household changes in healthful choices of HMS households that move during our sample period. Each year, Nielsen updates each HMS household’s zip code and county. HMS households change zip codes 17,956 times between 2004 and 2015, and they change counties 10,498 times. Restricting the sample to a balanced panel in which we observe households for at least four consecutive years including the year before the move, the year of the move, and two years after the move in the same new location, yields 3,305 cross-zip moves and 3,124 cross-county moves.

We test whether the Health Index is associated with the (time-invariant) sample average Health Index of groceries purchased in the household’s current geographic location, conditional on household fixed effects. By conditioning on household fixed effects, we isolate the grocery purchase changes associated with changes in neighborhood variables generated by moves.

Define $H_n$ as the estimated local Health Index of packaged groceries purchased in geographic area $n$, where $n$ will be either a zip code or a county. For county-level $H_n$, we use the same estimates mapped in Figure 1, and we use the same approach to estimate $H_n$ by zip code. For these regressions, $\mu_t$ represents year indicators, $\phi_i$ is a household fixed effect, and $X_{it}$ is again the vector of potentially time-varying household covariates described in Table 1. We estimate the following regression in household-by-year HMS data:

$$Y_{it} = \tau H_n + \beta X_{it} + \mu_t + \phi_i + \varepsilon_{it}. \quad (4)$$

Since the set of movers are not nationally representative, we do not use the HMS sample weights for this analysis.

Before estimating Equation (2), we first show graphical results of the event study. $B_{it}$ is an indicator for whether observation $it$ is part of a balanced sample around a move: we observe the household in two consecutive panel years in different locations, with another $y$ consecutive years before in the initial location and $y$ consecutive years after in the final location. $H_f$ and $H_o$, respectively, are the average Health Index of grocery purchases in the final and original locations, respectively, and we define $\Delta_{fo} \equiv H_f - H_o$.

We estimate the following regression in household-by-year data:
\[ Y_{it} = B_{it} \cdot (\tau_y \Delta_{fo} + \omega_y + \delta \Delta_{fo}) + \beta X_{it} + \mu_t + \phi_i + \varepsilon_{it}. \]  

(5)

We estimate \( \tau_y \) coefficients from year \( y \) to \( \overline{y} \), with \( y = -1 \) as the omitted category. \( \omega_y \) measure the average Health Index of movers in each year before and after moves, \( \delta \) measures the association between Health Index and the change in local environment in the year before the move \( (y = -1) \), and \( \tau_y \) measures the differences in that association for each other year in the event study window. The interaction with \( B_{it} \) means that we identify these coefficients using only the households in the balanced sample, although we include the full sample in the regression to more precisely pin down demographic associations \( (\beta) \) and year effects \( (\mu_t) \).

Figure 8 presents the estimated \( \tau_y \) coefficients and 95 percent confidence intervals. The move occurs during year 0 on the x-axis—that is, the household reports living in location \( o \) at the end of year \( -1 \) and location \( f \) by the end of year 0. We want to show as long a period as possible before and after moves, but the HMS sample thins out as we require increasingly long balanced samples around moves. For our figures, we thus estimate three different windows: \( y = 3 \) to \( \overline{y} = 1 \) to test for pre-move trends, \( y = 1 \) to \( \overline{y} = 2 \) to deliver precise short-term impacts, and \( y = 1 \) to \( \overline{y} = 3 \) to deliver a slightly longer-term estimate. We then present all three estimates on the figures.

Panel (a) of Figure 8 presents estimates for moves across counties, while Panel (b) presents estimates for moves across zip codes. In both panels, there is no pre-move trend in the association between Health Index and the local environment change \( \Delta_{fo} \), and no statistically significant changes in that association after the move. The 95 percent confidence intervals rule out post-move \( \tau_y \) coefficients of larger than about 0.05 to 0.15, depending on the specification. This means that when a household moves between areas where average healthy eating patterns differ by amount \( x \), we can reject that the household’s own eating patterns change by 0.05 to 0.15 within the next 1-3 years. Appendix Figure A10 presents these figures estimated without the demographic controls \( X_{it} \); the results are very similar.

A concern with our interpretation of \( \tau_y \) is that households may be moving for reasons that create endogeneity concerns. The parameters \( \omega_y \) should control for temporary changes in eating patterns due to the move itself or systematic differences between movers and non-movers. On the other hand, if some unobserved factor such as a better job or desire for a lifestyle change both causes healthier (or less healthy) eating and causes moves to a healthier place, the \( \tau \) coefficients may not capture the causal effect of “place effects.” In Appendix Figure 8 and Appendix Table A10, for example, we show that moving to healthier counties is statistically significantly associated with income increases, although moving to healthier zip codes is not.

We address this concern in two ways. First, we include controls for observable household demographics \( X_{it} \), which includes natural log of current-year income, natural log of years of education, indicators for each integer age from 23-90, an indicator for the presence of children, race indicators, a married indicator, employment status, weekly work hours, and household daily calorie need. This helps to control for observed changes in income, job responsibilities, household composition, and
marriage status that could generate endogeneity. For both the graphical estimates of Equation (5) and the regression tables from Equation (4), the $\tau$ coefficient estimates are very similar regardless of the inclusion of $X_{it}$.

Second, we assume that any remaining endogeneity would bias $\tau_y$ upward. This is the natural direction of bias for lifestyle changes that cause people to move to healthier places and also cause healthier eating, or salary increases that cause moves to higher-income (and healthier) places and also allow healthier eating. Under this direction-of-bias assumption, our $\hat{\tau}$ is an upper bound on the causal place effect, which biases against our finding of no place effects.

Table 4 presents estimates of Equation (4). Columns 1 and 2 consider cross-zip code moves, while columns 3 and 4 consider cross-county moves. Sample sizes are slightly smaller in columns 1-2 because $H_n$ is missing for zip codes with no stores. In all four columns, $\hat{\tau}$ is both statistically and economically insignificant. At the 5% significance level, we can reject values of $\tau$ greater than about 0.015 in columns 1-2, and about 0.09 in columns 3-4. Thus, moving to a place with $x$ units higher Health Index is associated with less than a 0.015$x$ to 0.09$x$ increase in a household’s Health Index within the next few years. Including controls for household demographics $X_{it}$ has very little impact on the results.

The lack of place effects on healthy eating contrasts with Bronnenberg, Dubé, and Gentzkow (2012)’s estimates of strong effects of the effects of local brand market shares on individual brand purchases. As an example of this brand preference effect, we estimate Equation (4) using county-level Coke market share for $H_n$, where “Coke market share” is Coke calories purchased/(Coke+Pepsi calories purchased). As shown in Appendix Table A9 we estimate a highly statistically significant $\hat{\tau} \approx 0.12$: moving to a county with (say) a 10 percentage point higher Coke market share is associated with about a 1.2 percentage point increase in the share of household Coke+Pepsi purchases that are Coke. Thus, using the same regressions with cross-county moves, we can rule out a place effect on healthy eating of anything more than about one-eighth the size of the place effect on Coke/Pepsi brand choice.

We now quantify the extent to which location explains the nutrition-income relationship in household shopping. The average household with income above $70,000 lives in a zip code (county) with a Health Index 0.13 (0.10) higher than households with incomes below $20,000. The upper bounds of the confidence intervals on $\tau$ for zip codes (counties) from Table 4 are about 0.015 (0.09), and the difference between the high and low-income Health Index is 0.29 standard deviations. Thus, in combination with the assumption that any endogeneity would bias $\tau$ upward relative to the causal effect of place, we conclude that place effects explain no more than $0.13 \times 0.015/0.29 \approx 0.65\%$ of the high- vs. low-income difference in Health Index using cross-zip code moves, and no more than $0.10 \times 0.09/0.29 \approx 3\%$ using cross-county moves.
V A Model of Nutrient Demand

We use a structural model of a household’s weekly grocery purchase decisions to quantify the magnitude of preferences as a driver for the nutrition-income relationship. Let $\mathcal{S}$ denote the set of stores considered by the household. We assume the household has full information about the prices and availability of products across all the stores in $\mathcal{S}$. In a given week, the household visits a subset $s \in \mathcal{P}(\mathcal{S})$ of the local stores and incurs shopping costs $\alpha d(s)$ where $\alpha$ is a dollar-denominated travel cost per hour and $d(s)$ is the total travel time.

Each week, the household jointly decides which stores to visit and what bundle of goods to purchase to determine its total consumption of calories and nutrients. Let $y = (y_1, ..., y_N)$ denote the quantities (measured in calories) of each of the $N$ food products (UPCs) stocked across all the stores with corresponding prices (per calorie) $p = (p_1, ..., p_N)$. Let $\Psi = (\Psi_1, ..., \Psi_N)$ denote the perceived qualities of each of the goods. Finally, let $x$ denote the composite good capturing all the other weekly expenditures with price normalized to $p_x = 1$ and perceived quality $\Psi_x = 1$.

Each of the $n = 1, ..., N$ products is characterized by $C$ nutrient characteristics $\{a_{n1}, ..., a_{nC}\}$. Define the $N \times C$ matrix $A = \begin{pmatrix} a_{11}, ..., a_{1C} \\ \vdots \\ a_{n1}, ..., a_{nC} \end{pmatrix}$, which measures the nutrient content (in kilograms) per calorie for each of the $C$ nutrients in each of the $N$ different products. The $C \times 1$ vector $z = A'y$ denotes the total nutrient consumption associated with the household’s bundle of calories.

Each week, the household solves the following utility-maximization problem:

$$\max_{s \in \mathcal{P}(\mathcal{S}), x, y} U(x, z, y; \Theta, \Psi) - \alpha d(s)$$

s.t.

$$\sum_{n=1}^{N} y_n p_n + x \leq I$$

(6)

to balance both its calories and nutrients purchased subject to its budget constraint, $I$, and the opportunity cost of the time spent “shopping”, $\alpha d(s)$. We assume the utility function $U(x, z, y; \Theta, \Psi)$ is continuous, increasing and strictly quasi-concave. $\Theta$ is a vector of taste parameters. Since $U$ is increasing, the household will spend its entire budget (the budget constraint will bind) and at least one good will always be consumed. We assume an interior quantity of the composite good is always consumed.

V.A Calorie Demand

Suppose the household visits the subset of stores $s \in \mathcal{P}(\mathcal{S})$. Without loss of generality, suppose we partition all the products into $n = 1, ..., N$ goods at the stores visited by the household and $n = N+$
1, ..., \(N\) goods in the stores not visited. Thus, \(y_n \geq 0\) (\(n = 1, ..., N\)) and \(y_n = 0\) (\(n = N + 1, ..., N\)). The optimal calorific consumption for each of the \(n = 1, ..., N\) purchased goods, \(y^*_n (p_s; \Theta, \Psi)\), satisfy the following system of first-order necessary conditions:

\[
\sum_{c=1}^{C} a_{nc} \frac{\partial U(x^*, z^*, y^*: \Theta, \Psi)}{\partial z_c} - \frac{\partial U(x^*, z^*, y^*: \Theta, \Psi)}{\partial y_n} p_n + \frac{\partial U(x^*, z^*, y^*: \Theta, \Psi)}{\partial y_n} y^*_n = 0, \quad n = 1, ..., N \tag{7}
\]

where \(p_s\) is a vector prices of all the goods in the set of considered stores. The household’s outside good expenditure is \(x^* = I - \sum_{n=1}^{N} y^*_n (p_s; \Theta, \Psi) p_n\).

We let \(y^*_s (p_s, \lambda_s; \Theta, \Psi)\) denote the vector of demands conditional on visiting the set of stores \(s\).

The conditional indirect utility associated with shopping in the set of stores \(s\) is

\[
v_s (p_s; \Theta, \Psi) = U \left( I - p_s' y^*_s (p_s; \Theta, \Psi), A' y^*_s (p_s; \Theta, \Psi), y^*_s (p_s; \Theta, \Psi) ; \Theta, \Psi \right) - \alpha d(s).
\]

The household’s optimal store choice problem can then be written as the following discrete choice problem:

\[
s^* (p; \Theta, \Psi) = \max_{s \in \mathcal{P}(S)} \{ v_s (p_s; \Theta, \Psi) \} \forall s \in \mathcal{P}(S) \tag{8}
\]

V.B A CES Model of Utility

To formulate a tractable model, we use the Dubois, Griffith, and Nevo (2014) utility framework with additively separable CES preferences over calories from each of \(J\) product categories and Cobb Douglas preferences over nutrients:

\[
U (x, z, y; \Theta, \Psi) = \sum_{j=1}^{J} \mu_j \log \left( \sum_{k=1}^{K_j} \Psi_{kj} \theta_j y_{kj} \right) + \sum_{c=1}^{C} \beta_c z_c + \gamma x.
\]

The \(N\) products (across all \(S\) stores) have been grouped into \(J\) product categories, such as carbonated soft drinks, bread and milk. Each category contains \(k = 1, ..., K_j\) products. The parameter \(\mu_j\) captures a household’s satiation rate over calories consumed in category \(j\). \(\theta_j\) determines the household’s satiation rate over calories consumed through product \(j\) in category \(J\). \(\Psi_{kj}\) allows for perceived product differentiation so that the household’s marginal benefit of calories can differ across brands within a category. \(\gamma\) represents the marginal utility of income spent on the outside good. Finally, \(\beta_c\) represents the marginal utility of consumption of nutrient \(c\).

For a given considered set of stores, \(s \in \mathcal{P}(S)\), only \(N \leq N\) of these will be purchased, and we let \(K_j\) denote the considered set of products sold in category \(j\) across the visited stores, where \(\sum_j K_j = N\).

We can re-write the first-order conditions, 7, as follows:

\[
\frac{\mu_j \theta_j}{\gamma} + \sum_{c=1}^{C} \frac{\beta_c}{\gamma} \sum_{k=1}^{K_j} a_{kjc} y^*_{jk} = \sum_{k=1}^{K_j} p_{kj} y^*_{kj}, \quad j = 1, ..., J \tag{9}
\]
where we have also summed over all the products in each product group $j$. Equation 9 describes a household’s optimal expenditure in category $j$ as a function of total nutrients consumed and category preferences as well as the shadow price $\lambda_s$.

VI Estimation and Results

VI.A Empirical Model

Let $h = 1, \ldots, H$ index households and $t = 1, \ldots, \tau_T$ index weeks in a year $T$. We can now write the panel version of our model as follows:

$$\sum_{t \in T} \mu_{htj} \theta_{htj} + \sum_{c=1}^{C} \frac{\beta_{hc}}{\gamma_h} \sum_{k=1}^{K_j} \sum_{t \in T} a_{kjc} y_{htjkt} = \sum_{t \in T} \sum_{k=1}^{K_j} p_{ktj} y_{htjkt}$$

(10)

where we have now aggregated category expenditures to the annual level. We use an annual aggregation to allow for more complex, higher-frequency correlations within a household’s shopping history (e.g. stock-piling, habits etc). Equation 10 relates a household’s optimal annual expenditure in a given category to its average category preferences and total annual calories consumed. $\sum_{t \in T} \frac{\mu_{htj} \theta_{htj}}{\gamma_h}$ corresponds to household $h$’s annual expenditure in product group $j$, absent the nutritional value of products in product group $j$. The term $\sum_{c=1}^{C} \frac{\beta_{hc}}{\gamma_h} \sum_{k=1}^{K_j} \sum_{t \in T} a_{kjc} y_{htjkt}$ captures household’s $h$’s additional expenditures in the category due to the desirability of the nutrient contents of the products. The model potentially allows for considerable heterogeneity across households and time.

In principle, we could use 10 to derive the following empirical specification:

$$m_{hetj} = \delta_{htj} + \beta_c Z_{htj} + \epsilon_{htj}$$

where $m_{hetj}$ is a household’s annual expenditure in category $j$ and $Z_{htj}$ is annual nutrient purchases and $\epsilon_{htj}$ contains household-specific annual deviations from mean tastes. The problem with this specification is that the calorie demand in a category appears on both the left and right hand sides.

It will be useful for estimation to rewrite equation 10 by solving for total calories purchased by household $h$ in product group $j$ in year $T$. Define total calories purchased by household $h$ in product group $j$ in year $T$ as: $Y_{htj} = \sum_{t \in T} \sum_{k \in C} y_{htjkt}$. Equation 10 can now be written as:

$$\sum_{t \in T} \frac{\mu_{htj} \theta_{htj}}{\gamma_h} + \sum_{c=1}^{C} \frac{\beta_{hc}}{\gamma_h} \tilde{a}_{htj} Y_{htj} = \tilde{p}_{htj} Y_{htj},$$

(11)

where $\tilde{p}_{htj}$ is the calorie weighted average price paid per calorie by household $h$ in product group $j$ in year $T$. Similarly, $\tilde{a}_{htj}$ is the calorie weighted average amount of nutrient $c$ per calorie in products purchased by household $h$ in product group $j$ in year $T$. Equation 11 can now be solved
for total calories, \( Y_{hjT} \):

\[
Y_{hjT} = \frac{\sum_{t \in T} \frac{\mu_{hjt} \theta_{hjt}}{\gamma_h}}{\tilde{p}_{hjT} - \sum_{c=1}^C \frac{\beta_{hc}}{\gamma_h} \tilde{a}_{hTjc}}.
\]  

(12)

Finally, taking logs of both sides yields:

\[
\log(Y_{hjT}) = \log\left(\sum_{t \in T} \frac{\mu_{hjt} \theta_{hjt}}{\gamma_h}\right) - \log(\tilde{p}_{hjT} - \sum_{c=1}^C \frac{\beta_{hc}}{\gamma_h} \tilde{a}_{hTjc}).
\]  

(13)

One of the most desirable properties of the model is that it allows for estimation of nutrient and product group preferences from data that have been aggregated to the annual product group level. This allows us to focus on estimation of preferences for nutrients without dealing with the parameters driving preferences at the UPC level or the dynamics in purchase behavior at the weekly level.

We base the estimation of household’s preferences for product groups and nutrients on Equation 13. We also allow for heterogenous preferences across households of different income types. Household \( h \)'s type is denoted \( b(h) \). We now define the household \( h \)'s preference for product group \( j \) in year \( T \) living in zip3 \( \text{zip}(h) \) by a product group fixed effect and zip3 fixed effects for households of type \( b(h) \), and the deviation from their sum:

\[
\log(\sum_{t \in T} \frac{\mu_{hjt} \theta_{hjt}}{\gamma_{ht}}) = \delta_{b(h)j} + \delta_{b(h)\text{zip}(h)} + \tilde{\delta}_{hjT}.
\]  

(14)

We also allow for one of the product nutrients to be unobserved to the econometrician and to vary across household types. Let the first \( c = 1, \ldots, C - 1 \) nutrients be observed and the \( C^{th} \) nutrient be unobserved. Preferences for nutrients, \( \beta \), and the outside good, \( \gamma \), will vary by household:

\[
\sum_{c=1}^C \frac{\beta_{b(h)c}}{\gamma_{b(h)t}} \tilde{a}_{hTjc} = \sum_{c=1}^{C-1} \tilde{\beta}_{b(h)c} \tilde{a}_{hTjc} + \tilde{\alpha}_{b(h)C}
\]  

(15)

\[
\tilde{\beta}_{b(h)c} = \frac{\beta_{b(h)c}}{\gamma_{b(h)}},
\]  

\[
\tilde{\alpha}_{b(h)C} = \frac{\beta_{b(h)C}}{\gamma_{b(h)}} \tilde{a}_C.
\]  

(16)

Plugging equations (14) and 15 into equation 13 gives:

\[
\log(Y_{hjT}) = F\left(\tilde{p}_{hjT}, \tilde{a}_{hTjc}; \tilde{\alpha}_{b(h)C}, \tilde{\beta}_{b(h)}\right) + \delta_{b(h)j} + \delta_{b(h)\text{zip}(h)} + \tilde{\delta}_{hjT}
\]  

(17)

where

\[
F\left(\tilde{p}_{hjT}, \tilde{a}_{hTjc}; \tilde{\alpha}_{b(h)C}, \tilde{\beta}_{b(h)}\right) = -\log\left(\tilde{p}_{hjT} - \sum_{c=1}^{C-1} \tilde{\beta}_{b(h)c} \tilde{a}_{hTjc} + \tilde{\alpha}_{b(h)C}\right).
\]
VI.B Identification

Applying non-linear least squares to equation 17 would not generate consistent estimates of our model parameters. There are several potential sources of endogeneity that still need to be addressed.

Our key identifying assumption is that households’ idiosyncratic preferences for a given product group, $\tilde{\delta}_{hjT}$ are uncorrelated with nutrient amounts $c$, $\tilde{a}_{hTjc}$. This assumption could potentially be violated if, for example, households purchased salty foods because they prefer product groups that have longer shelf lives. We will not be able to separate out how much of a preference for salty foods is due to a true nutrition value of sodium versus other impacts of sodium on the product. We identify the value of nutrients from a nutrition standpoint along with all the other impacts nutrients have on the desirability of food (taste, shelf life, ease of preparation.) It is unlikely that supply-side instrumental variables could solve this problem due to the challenge in creating variation only in the nutrition value of a product, without impacting taste or shelf life.

A related identifying assumption is that households’ idiosyncratic perceptions for each individual product $kj$’s quality, $\Psi_{hkj}$, is uncorrelated with its purchase intensity (or satiation) for the category, $\mu_{hjt}$ or its purchase intensity (or satiation) from specific products in the category, $\theta_{hjt}$. The quality term $\Psi_{hkj}$ drops out of the overall total category expenditure decision, equation 9. However, $\Psi_{hkj}$ influences the quantities purchased for each of the individual products within the category, $y_{hkjt}$, and hence the calorie-weighted average amount of each nutrient per calorie in the product group. Formally, we can then assume:

$$E\left((\tilde{\delta}_{hjT} + \delta_{(h)j})\tilde{a}_{hjT}\right) = 0.$$

VI.C Price Endogeneity

A more worrisome source of endogeneity arises from the potential correlation between a household’s preferences for the product category, $\tilde{\delta}_{hjT}$, and prices, $\tilde{p}_{hjT}$:

$$E\left(\tilde{\delta}_{hjT}\tilde{p}_{hjT}\right) \neq 0. \tag{18}$$

Price endogeneity could arise from both the demand and supply sides. On the demand side, recall the household’s store-choice problem in equation 8. The household’s conditional indirect utility associated with the set of considered stores depends on its preferences and its shopping costs. To the extent that unobserved preferences, $\delta_{hjT}$, or unobserved shopping costs, $\alpha$, correlate with the set of stores and prices they may not have access to the lowest prices per calorie. The store choices introduce self-selection into the set of prices facing a given household. Price endogeneity can also arise on the supply side if firms systematically charge different prices in different geographic markets. In our data, we observe variation across geographic markets in the prices charged for comparable goods.

To correct for both types of price endogeneity, we instrument the prices faced by household $h$
for product category \( j \) in period \( t \), \( \tilde{p}_{hjT} \). The underlying intuition for our instrument is that large supermarket chains differ in their sourcing and distribution costs across products, giving different chains heterogeneous comparative advantages in supplying certain types of food products. Using variation across zip3s in presence of each chain shifts which product groups are particularly cheap in each zip3. To illustrate, consider a simple example in which there are two types of foods – apples and pizza – and two grocery chains – Safeway and Shaws. Suppose Safeway is able to source pizza cheaply, while Shaws can source apples cheaply. Then, cities dominated by Safeway will face relatively low prices for pizza and cities dominated by Shaws will face relatively low prices for apples.

We construct our instrument as follows. For chain \( c \) in market \( m \) during time period \( t \), let \( \log(p_{k,c,-m,t}) \) denote the average log price of a given product (at the UPC level) \( k \) in stores from the same chain but in all zip3s excluding market \( m \), denoted by \(-m\). Let \( \log(p_{k,-m,t}) \) denote the national average log price of the UPC \( k \) in period \( t \) in all markets excluding \( m \). We exclude market \( m \) to ensure that the IV reflects a chain’s comparative advantages in supplying product \( k \) based on other markets. We measure chain \( c \)’s cost advantage in supplying UPC \( k \) relative to the national average as follows: \( \Delta \log(p_{k,c,-m,t}) = \log(p_{k,c,-m,t}) - \log(p_{k,t}) \). We then construct a product-group-level instrument for product group \( j \) in market \( m \) during period \( t \) as follows:

\[
\Delta \log(\bar{p}_{j,m,t}) = \sum_{k \in j,c \in m} \frac{1}{N_{jct} \times N_{kt} \times N_{cmt}} \sum_{k \in j,c \in m} N_{cmt} \times N_{jct} \times N_{kt} \times \log(\Delta p_{k,c,-m,t}).
\]

Thus, we weight each price deviation in product group \( j \) by the product group’s nationwide total units sold, \( N_{kt} \), total nationwide revenues at chain \( c \), and total number of chain \( c \) establishments in market \( m \) and year \( t \), \( N_{cmt} \). Our identifying assumption is:

\[
E\left( \delta_{hjtm} \Delta \log(p_{j,m(h),t}) \right) = 0. \tag{19}
\]

The variation in our instrument reflects geographic variation in the presence of each chain across zip3 markets. Since our model includes product group and zip3 fixed effects, our instrument relies on variation in the relative prices across product groups. The key assumption is that a chain with a comparative pricing advantage in supplying product group \( j \) does not base its entry decisions on a given market’s tastes for \( j \). We cannot explicitly rule out such an entry pattern. However, the geographic network of each chain is quite diverse and it is unlikely that the match between a market’s product group preferences and the chains’ relative pricing advantages plays a large role in entry decisions.

To assess the power of our instrument, Figure A12 presents the binscatter relationship between log prices and our instrument, controlling for zip3 and product group fixed effects. The plot shows a robust linear relationship, and Table A10 shows an F-stat of at least 83 across all four income groups.
VI.D GMM Estimation

We make the following additional assumptions to allow for heterogeneity across income groups in the empirical model:

\[
\begin{bmatrix}
\delta_h \\
\bar{\beta}_h
\end{bmatrix} =
\begin{cases}
\begin{bmatrix}
\delta_1 \\
\bar{\beta}_1
\end{bmatrix}, & \text{if Income}_h \leq Q_{25} (\text{Income}) \\
\begin{bmatrix}
\delta_2 \\
\bar{\beta}_2
\end{bmatrix}, & \text{if Income}_h \in (Q_{25} (\text{Income}), Q_{50} (\text{Income})) \\
\begin{bmatrix}
\delta_3 \\
\bar{\beta}_3
\end{bmatrix}, & \text{if Income}_h \in (Q_{50} (\text{Income}), Q_{75} (\text{Income})) \\
\begin{bmatrix}
\delta_4 \\
\bar{\beta}_4
\end{bmatrix}, & \text{if Income}_h > Q_{75} (\text{Income}).
\end{cases}
\]

The model estimation will rely on the following set of \((C + 1 + J + M)\) identifying moments:

\[
\begin{align*}
E\left( (\hat{\delta}_{hjT} + \delta_{hj})\tilde{a}_{hjTc} \right) & = 0, \quad c = 1, ..., C \\
E\left( \hat{\delta}_{hjT} \Delta\tilde{p}_{j,T,m(h)} \right) & = 0 \\
E\left( \hat{\delta}_{hjT} D_{hjT} \right) & = 0, \quad j = 1, ..., J \\
E\left( \hat{\delta}_{hjT} D_{zip(h)} \right) & = 0, \quad zip(h) = 1, ..., M
\end{align*}
\]

Let \(D_{hjT}\) be a \(J \times 1\) vector of dummy variables indicating the choice of product group \(J\) by household \(h\) on trip \(T\). Let \(D_{zip(h)}\) be an \(M \times 1\) vector of dummy variables indicating the zip3 of household \(h\). We construct the sample analog of the moment conditions from (20) in the following \((C + 1 + J + M) \times 1\) vector:

\[
\left[ \tilde{a}_{hjT}(\hat{\delta}_{hjT} + D_{hjT}\delta_h) \right] \\
\Delta\tilde{p}_{j,T,m(h)}\hat{\delta}_{hjT} \\
D_{hjT}\tilde{\delta}_{hjT} \\
D_{zip(h)}\tilde{\delta}_{hjT}
\]
We can then write the GMM estimator as:

\[
\left( \hat{\delta}, \hat{\delta}_{zip}, \hat{\beta} \right) = \arg\min_{\delta, \delta_{zip}, \tilde{\beta}} \left( \frac{1}{JHT} \sum_h \sum_T \sum_j g_{hjT} \left( \delta_h, \delta_{zip,h}, \tilde{\beta}_h \right) \right) \mathbf{W} \left( \frac{1}{JHT} \sum_h \sum_T \sum_j g_{hjT} \left( \delta_h, \delta_{zip,h}, \tilde{\beta}_h \right) \right)'
\]

where \( W \) is a \((C + 1 + J + M) \times (C + 1 + J + M)\) weight matrix. Details regarding the GMM estimator and its standard errors are in Appendix D.

VI.E Estimation Results

Table 5 reports the estimated preferences for product nutrients across the four income groups (i.e. across income quartiles). These estimates have been normalized to represent the total kilograms of a given nutrient required to generate the same utility as a kilogram of carbohydrate. This normalization removes differences in the marginal utility of a dollar \( \gamma_{b(i)} \) across the income group\(^{21}\)

Our point estimates indicate a striking monotonicity in the preferences for fat, saturated fat, fiber, protein, sugar, sodium, fruit and vegetables in the level of household income. Higher income groups seem to exhibit more desire for healthy nutrients. Higher income groups have a stronger desire for fiber, protein, fruit and vegetables. Higher income groups also have a weaker taste for fat, saturated fats, sodium and sugar, leading them to make more healthy choices. While fat is viewed as undesirable (relative to carbohydrate) in all income groups, it is nearly 20 times more undesirable to the highest-income group than the lowest. Similarly, relative to carbohydrate, the highest-income group values fruit nearly three times as much as the lowest-income group, and vegetables nearly twice as much.\(^{22}\) Interestingly, the four income groups have almost identical preferences for cholesterol. The unobserved nutrient is also increasinlgy desirable for higher income groups, suggesting unobserved quality of the products purchases by higher income households is higher. Overall, we observe a strong relationship between income and preferences for healthy nutrients.

VII Nutrition Choices and Preferences

VII.A Decomposing nutrition choices into supply and demand factors

Using the model estimates from section VI.E, we decompose how much of the observed nutrition differences observed across the income groups are due to supply-side factors (prices and availability of product nutrients) versus demand-side factors (product group and nutrient preferences). Since

\(^{21}\)One restriction of this model is that the Cobb-Douglas utility function restricts the price elasticity to one at the product group level. We test this assumption by estimating a reduced-form price elasticity using 2SLS. These estimates are in Appendix Table A10. Consistent with our model, all four income groups exhibit price elasticities of product group demand of one. Note that the product group level elasticities are likely very different from the individual product (UPC) level price elasticities which may be very heterogeneous across income groups.

\(^{22}\)These estimates for the value of fruit and vegetables measures their respective nutrient desirability other than through our standard observed macronutrients. This reflects the value of the vitamins found in fruit and vegetables.
our model is estimated at the product group level, our counterfactuals only allow households to re-optimize their calorie demand across product groups. We do not analyze how households would change their relative quantities across UPCs within product groups.

For a given set of prices for each product group ($\tilde{p}_j$), nutrients for each product group $j$ ($\tilde{a}_{jc_1}, ..., \tilde{a}_{jc_o}, \tilde{\alpha}_{cu}$), product group preferences ($\tilde{\delta}_1, ..., \tilde{\delta}_j$), and nutrient preferences ($\tilde{\beta}_1, ..., \tilde{\beta}_{C_o}$), we can calculate the total calories consumed within each product group, $\hat{Y}_j$:

$$\hat{Y}_j = \frac{\exp(\tilde{\delta}_j)}{\tilde{p}_j - \sum_{c=1}^{C_o} \tilde{\beta}_c \tilde{a}_{jc_c} - \tilde{\alpha}_{cu}}.$$ (22)

To evaluate the healthfulness of this bundle of goods, we use the same health health index as in Section II.C, where $w_c$ are the health index weights on each nutrient:

$$\hat{H} = \sum_j \sum_c w_c \tilde{a}_{jc} \hat{Y}_j.$$ (23)

Figure 9 displays each income group’s base health index level by computing the health index at the average prices and nutrient levels available across each household-year in the group.

We begin with the supply side. Our first counterfactual explores the role of product group prices in healthy eating differences across the income groups. For each income group, we set each product group’s prices to the levels observed in the highest income group. Thus, for each income group $b(i)$, we calculate the product group calorie demand as:

$$\hat{Y}_{b(i)} = \frac{\exp(\tilde{\delta}_{b(i)j})}{\tilde{p}_{4j} - \sum_{c=1}^{C_o} \tilde{\beta}_{b(i)c} \tilde{a}_{4jc_c} - \tilde{\alpha}_{4cu}}.$$ (24)

We then compute the corresponding health index by substituting the quantities in (24) into the health index (23). Figure 9 indicates that prices do not appear to explain much of the gap in healthfulness across income groups. Recall from section III.C that healthful foods are in fact slightly more expensive in the high-income groups. So the equalization of prices actually increases the gap between the highest and lowest income groups slightly by 2% relative to base.

Our second counterfactual explores the food desert hypothesis by setting the product group nutrient levels in each product group equal to those observed in the highest income group. We make this change in addition to equating the price level across income groups. Once again, we recompute each income group’s total calorie demand in product group $j$ as follows:

$$\hat{Y}_{b(i)j} = \frac{\exp(\tilde{\delta}_{b(i)j})}{\tilde{p}_{4j} - \sum_{c=1}^{C_o} \tilde{\beta}_{b(i)c} \tilde{a}_{4jc_c} - \tilde{\alpha}_{4cu}}.$$ (25)

Figure 9 now shows that the health index gap between the highest and lowest income groups declines by 8.9% relative to base. The first two counterfactuals confirm our findings from section III.B. We find that by changing the availability of nutrients available in the lowest-income group
to be the same as that of the highest income group does reduce the healthfulness gap, but by less than 10%. Accordingly, the so-called food desert hypothesis does not appear to explain very much of the nutrition-income relationship.

We now explore the role of demand-side differences. In addition to the changes in prices and nutrient availability, we also set the nutrient preferences in each income group to those of the highest-income group, generating total calorie demand:

$$\hat{Y}_{b(i)j} = \exp(\delta) \frac{\tilde{p}_j - \sum_{c=1}^{C_o} \tilde{\beta}_{4c} \tilde{a}_{4jc} - \tilde{\alpha}_{4ca}}{}.$$  (26)

Figure 9 shows that equalizing the nutrient preferences closes most of the gap in healthy eating. While the bundles chosen by the highest-income group are still healthier than those of the lower income groups, the gap between the highest and lowest income groups declines by 84% relative to base.

Finally, in addition to prices, nutrient availability and nutrient preferences, we also set the product group preferences preferences to those of the highest income group, generating total calorie demand:

$$\hat{Y}_{b(i)j} = \exp(\delta_{ij}) \frac{\tilde{p}_j - \sum_{c=1}^{C_o} \tilde{\beta}_{4c} \tilde{a}_{4jc} - \tilde{\alpha}_{4ca}}{}.$$  (27)

By construction, this last counterfactual mechanically equalizes the observed purchases across each of the income groups, as seen in Figure 9.

Figure 10 summarizes the decomposition (in percentage terms) of the healthy eating gap between the highest and lowest income groups across each of these four different factors on the supply and demand sides. Overall, prices explain about -2.3%, product nutrients explain 11.1%, product group preferences explain 16.2% and nutrient preferences explain 75%. Most notably, over 90% of the healthy eating gap is due to demand-side factors related to preferences.

### VII.B Education and knowledge

In the previous section, we parsed out the effect of nutrient supply and nutrient prices on healthful choices. We now study the extent to which other observable household characteristics might explain part of the relationship between income and healthy grocery purchases.

To control for supply, we hold the prices for each product group $j$, $\tilde{p}_j$, and nutrients for each product group $j$, $(a_{jc_1}, ..., a_{jc_o}, \tilde{\alpha}_{ca})$, fixed at their mean levels across all household trips. For a household $i$, we then compute its total calorie demand in product group $j$ as follows:

$$\hat{Y}_{ij} = \frac{\exp(\delta_{0ij} + \delta_{0(i),zip(i)} + \tilde{\delta}_{ijT})}{\tilde{p}_j - \sum_{c=1}^{C_o} \beta_{0c} \tilde{a}_{0jc} - \alpha_{0ca}}.$$  (28)

so that heterogeneity in healthfulness within an income group is captured by the unobserved com-
ponent of demand, $\delta_{ijT}$. Finally, we compute each household’s health index:

$$\hat{H}_i = \sum_j \sum_c w_c \tilde{a}_{jc} \hat{Y}_{ij}. \quad (29)$$

To study the role of household traits, we regress each household $i$’s Health Index (29) on the natural logarithm of income, $\ln I_{it}$, and household demographics $X_i$:

$$\hat{H}_i = \alpha \ln I_i + \beta X_i + \epsilon_i. \quad (30)$$

We estimate the extent to which each covariate $X_{ik}$ explains $\alpha$, the nutrition-income relationship. Following Gelbach (2016), we estimate the full model in Equation (30), which gives coefficients $\hat{\beta}_k$. We then estimate the auxiliary regression $\ln I_i = \Gamma X_{ki} + \epsilon_i$ to determine $\hat{\Gamma}$, the covariance between log income and the variable $X_{ik}$. We can then estimate variable $X_{ik}$’s contribution to the nutrition-income relationship by using the omitted variable bias formula: $\hat{\delta}_k = \hat{\Gamma}_k \hat{\beta}_k$. If $\hat{\alpha}$ denotes the estimated “unconditional” relationship between $Y$ and $\ln I$ (without conditioning on $X$), variable $X_{ik}$’s estimated contribution to the nutrition-income relationship is:

$$\tilde{\delta}_k = \frac{\hat{\Gamma}_k \hat{\beta}_k}{\hat{\alpha}}. \quad (31)$$

As Gelbach (2016) points out, the approach is equivalent to determining the omitted variable bias in $\alpha$ from omitting $X_k$.

Table 6 presents results. Panel A (“Unconditional Estimates”) presents the estimate of $\hat{\alpha}$, estimated from Equation (30) but excluding the $X$ covariates. Panel B (“Full Model”) presents estimates of the full Equation (30). Standard errors are clustered by household and observations are weighted for national representativeness.

Column 1 presents estimates defining $X$ using observed household demographics from the Nielsen HMS sample. The unconditional estimates show that a 100 log point increase in household income is associated with a 0.13 increase in the Health Index. Controlling for education, age, children, race, and census tract indicators attenuates the coefficient to 0.09. We present each variable’s $\tilde{\delta}_k$ below the standard error.

Education explains 27 percent of the nutrition-income relationship, a result that is qualitatively consistent with results in Handbury, Rahkovsky, and Schnell (2015). This finding could suggest that health knowledge, information, and cognition might be important (see for instance Cutler and Lleras-Muney, 2010). Weekly work hours also explains a very large and significant share of the relationship; though employment offsets almost half the $\delta_k$ for work. Marriage, race and presence of children have significant, but relatively small effects.

To further explore the potential importance of health knowledge and information, Column 2 adds the survey variables from the Bronnenberg et al. (2012) and Bronnenberg et al. (2015) to
Since the surveys were only collected in 2010, only a subset of the sample from column 1 can be matched. Column 2 shows that while working in a health-related occupation (which includes doctors, dentists, nurses, nutritionists, physical therapists, and others) is not associated with the Health Index, the food knowledge score is statistically significant and can explain 1% of the nutrition-income gradient. While this is a small effect, our measure of food knowledge is quite coarse and could likely be improved with better survey data directly measuring knowledge about nutrition.

VIII Conclusions

We study how and why healthful eating varies by income in the United States. The public health literature has documented that lower-income neighborhoods suffer from lower availability of healthful groceries and that lower-income households eat less healthfully. Some researchers and policymakers have concluded that food deserts, areas without supermarkets and, hence, with less access to healthful foods and charging potentially higher prices, cause lower-income households to purchase less healthful food. We find that despite the differences in local supply, low-income households still purchase the great majority of their groceries from grocery stores, supercenters, and club stores, which carry more produce and generally more healthful items than convenience stores, drug stores, and mass merchants. The notion of a food desert may therefore be misleading as it is typically based on a market definition that under-states households’ willingness-to-travel. Thus, when a new supermarket enters nearby, this may benefit households by reducing travel costs, but it does not meaningfully change their choice sets or the healthfulness of their purchases.

We therefore turn to the demand side, estimating household nutrient preferences. Our structural estimates of preferences indicate striking differences in tastes between low and high income households: higher-income groups have healthier nutrient preferences. We observe a monotonically increasing preference for healthful nutrients across income groups. We also find a monotonically decreasing preference for unhealthful nutrients across income groups.

We then use our structural estimates to analyze the relative importance of supply-side versus demand-side factors. On the supply side, we find that resolving the so-called food desert hypothesis (i.e. prices and nutrient availability) only eliminates at most 11% of the healthful eating gap. On the demand side, we find that the association between income and preferences for specific product groups and for nutrients explains over 90% of the gap in healthful eating. In an exploratory analysis, we find that education is an important driver of the association between income and differences in preferences for healthful eating. We find some evidence that knowledge about foods may a mechanism through which education and income correlated with healthful eating. These findings

---

Recall that in these surveys, panelists were asked to identify the most common additive to table salt (correct answer: iodine), the scientific name for baking soda (correct answer: sodium bicarbonate), and the most common ingredient of granulated sugar (correct answer: sucrose). The “Food knowledge” variable is the share of these questions answered correctly, normalized to mean 0, standard deviation 1.
suggest that policies geared towards nutrition education may be more effective than subsidies and grants geared towards food retailers.
References


# Tables

Table 1: **Descriptive Statistics**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard deviation</td>
<td>Grocery</td>
</tr>
<tr>
<td>Household income ($000s)</td>
<td>62.2</td>
<td>45.4</td>
<td>Large grocery (&gt;50 employees)</td>
</tr>
<tr>
<td>Years education</td>
<td>13.9</td>
<td>2.06</td>
<td>Supercenters/club stores</td>
</tr>
<tr>
<td>Age</td>
<td>52.2</td>
<td>14.4</td>
<td>Drug stores</td>
</tr>
<tr>
<td>Have children</td>
<td>0.33</td>
<td>0.47</td>
<td>Convenience stores</td>
</tr>
<tr>
<td>White</td>
<td>0.77</td>
<td>0.42</td>
<td>Meat/fish/produce stores</td>
</tr>
<tr>
<td>Black</td>
<td>0.12</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>0.49</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>0.61</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Weekly work hours</td>
<td>21.0</td>
<td>15.0</td>
<td></td>
</tr>
<tr>
<td>Household daily calorie need</td>
<td>5282</td>
<td>2987</td>
<td></td>
</tr>
<tr>
<td>Health occupation</td>
<td>0.055</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Food knowledge</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Homescan data include 668,844 household-by-year observations for 2004-2015 and are weighted for national representativeness. Health occupation and Food knowledge are from Homescan add-on surveys carried out by Bronnenberg et al. (2012) and Bronnenberg et al. (2015). Health occupation is an indicator for whether the respondent’s occupation is classified as “health care practitioners and technical,” which includes registered nurses, health technicians, doctors and dentists, pharmacists, physical therapists, dietitians and nutritionists, and related occupations. Food knowledge is the share of three questions answered correctly (identify the most common additive to table salt (correct answer: iodine), the scientific name for baking soda (correct answer: sodium bicarbonate), and the most common ingredient of granulated sugar (correct answer: sucrose)), normalized to mean zero, standard deviation one. Zip code establishment counts are from Zip Code Business Patterns data for 2004-2015, with 470,229 zip code-by-year observations. UPC characteristics are for all 1.57 million UPCs that ever appear in the Nielsen Homescan or RMS data.
### Table 2: Correlates of Health Index of UPCs Available at RMS Stores

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>(1) Count of Produce UPCs Offered</th>
<th>(2) Mean Health Index of UPCs Offered</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(Zip median income)</td>
<td>389.7 (8.85)***</td>
<td>-1.16 (5.30)***</td>
</tr>
<tr>
<td>ln(ACV)</td>
<td>353.4 (1.10)***</td>
<td>0.29</td>
</tr>
<tr>
<td>1(Large grocery)</td>
<td>667.0 (34.7)***</td>
<td>-0.69 (0.059)***</td>
</tr>
<tr>
<td>1(Small grocery)</td>
<td>249.1 (34.1)***</td>
<td>-0.87 (0.055)***</td>
</tr>
<tr>
<td>1(Supercenter/club)</td>
<td>75.7 (36.5)***</td>
<td>-0.91 (0.060)***</td>
</tr>
<tr>
<td>1(Convenience store)</td>
<td>-916.2 (34.1)***</td>
<td>-1.79 (0.058)***</td>
</tr>
<tr>
<td>1(Drug store)</td>
<td>-856.8 (34.0)***</td>
<td>-1.89 (0.058)***</td>
</tr>
<tr>
<td>1(Other mass merchant)</td>
<td>-781.7 (33.1)***</td>
<td>-1.78 (0.057)***</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.043 0.80</td>
<td>0.95 0.035</td>
</tr>
</tbody>
</table>

Notes: This table uses 2006-2015 Nielsen RMS data at the store-by-year level. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. 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. There is no omitted store type in columns 3 and 6. Robust standard errors, clustered by zip code, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
Table 3: Effects of Supermarket Entry

<table>
<thead>
<tr>
<th>Sample:</th>
<th>Full Sample</th>
<th>Income &lt; $25,000</th>
<th>“Food Desert” Zip Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Effects on Expenditure Shares</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expenditure shares at store type:</td>
<td>Entrants</td>
<td>Grocery/Super/Club</td>
<td>Entrants</td>
</tr>
<tr>
<td>Post entry: 0-10 minutes</td>
<td>3.781</td>
<td>0.634</td>
<td>4.737</td>
</tr>
<tr>
<td>Post entry: 10-15 minutes</td>
<td>1.075</td>
<td>0.184</td>
<td>1.019</td>
</tr>
<tr>
<td>Observations</td>
<td>2,606,413</td>
<td>2,606,413</td>
<td>400,690</td>
</tr>
<tr>
<td>Dependent var. mean</td>
<td>19</td>
<td>88</td>
<td>21</td>
</tr>
<tr>
<td><strong>Panel B: Effects on Health Index</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post entry: 0-10 minutes</td>
<td>0.00534</td>
<td>0.0102</td>
<td>-0.00683</td>
</tr>
<tr>
<td>Post entry: 10-15 minutes</td>
<td>0.00282</td>
<td>0.0114</td>
<td>0.0264</td>
</tr>
<tr>
<td>Observations</td>
<td>2,606,413</td>
<td>400,690</td>
<td>627,000</td>
</tr>
</tbody>
</table>

Notes: This table uses 2004-2015 Nielsen Homescan data at the household-by-quarter level. “Food desert” zip codes are those with no grocery stores with 50 or more employees, supercenters, or club stores in 2003. Expenditure shares are in units of percentage points. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Reported independent variables are indicators for whether a specific retailer has 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, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need), 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, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
Table 4: **Association of Health Index with Local Area Health Index Using Movers**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zip code average Health Index</td>
<td>-0.00575</td>
<td>-0.00375</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0103)</td>
<td>(0.0102)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>County average Health Index</td>
<td></td>
<td></td>
<td>0.0302</td>
<td>0.0327</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.0295)</td>
<td>(0.0292)</td>
</tr>
<tr>
<td>Household demographics</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>655,977</td>
<td>655,977</td>
<td>664,156</td>
<td>664,156</td>
</tr>
<tr>
<td>95% confidence interval upper bound</td>
<td>0.014</td>
<td>0.016</td>
<td>0.088</td>
<td>0.090</td>
</tr>
</tbody>
</table>

Notes: This table uses 2004-2015 Nielsen Homescan data at the household-by-year level. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Household demographics are natural log of income, natural log of years of education, age indicators, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need. All regressions also control for year indicators and household fixed effects. Observations are not weighted for national representativeness. Robust standard errors, clustered by household, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
<table>
<thead>
<tr>
<th>Income</th>
<th>Carbs</th>
<th>Fat</th>
<th>Sat. Fat</th>
<th>Fiber</th>
<th>Protein</th>
<th>Sugar</th>
<th>Sodium</th>
<th>Cholesterol</th>
<th>Fruit</th>
<th>Veg</th>
<th>Unobs. Nutrient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inc ≤ 25K</td>
<td>-</td>
<td>15.91</td>
<td>9.90</td>
<td>5.96</td>
<td>1.89</td>
<td>-29.78</td>
<td>-13.45</td>
<td>-30.03</td>
<td>0.38</td>
<td>0.39</td>
<td>-0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.11]</td>
<td>[0.17]</td>
<td>[0.28]</td>
<td>[0.12]</td>
<td>[1.35]</td>
<td>[3.24]</td>
<td>[0.02]</td>
<td>[0.01]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25K &lt; Inc ≤ 50K</td>
<td>-</td>
<td>13.70</td>
<td>10.19</td>
<td>6.63</td>
<td>1.31</td>
<td>-29.93</td>
<td>-13.45</td>
<td>-30.03</td>
<td>0.66</td>
<td>0.54</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.03]</td>
<td>[0.04]</td>
<td>[0.08]</td>
<td>[0.14]</td>
<td>[0.98]</td>
<td>[2.40]</td>
<td>[0.01]</td>
<td>[0.003]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50K &lt; Inc ≤ 75K</td>
<td>-</td>
<td>13.03</td>
<td>10.29</td>
<td>6.83</td>
<td>1.17</td>
<td>-29.97</td>
<td>-13.45</td>
<td>-30.03</td>
<td>0.85</td>
<td>0.63</td>
<td>-0.044</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.04]</td>
<td>[0.05]</td>
<td>[0.17]</td>
<td>[0.17]</td>
<td>[0.12]</td>
<td>[3.67]</td>
<td>[0.01]</td>
<td>[0.002]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75K &lt; Inc</td>
<td>-</td>
<td>12.50</td>
<td>10.38</td>
<td>7.32</td>
<td>1.06</td>
<td>-30.03</td>
<td>-13.45</td>
<td>-30.03</td>
<td>1.14</td>
<td>0.75</td>
<td>0.0003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.04]</td>
<td>[0.04]</td>
<td>[0.13]</td>
<td>[0.18]</td>
<td>[1.38]</td>
<td>[4.55]</td>
<td>[0.01]</td>
<td>[0.004]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Standard errors are clustered by household. Magnitudes represent kilograms of the nutrient which offers the same utility as a kilogram of carbohydrates. A given nutrient can enter through multiple preference parameters, such as saturated fat is both valued as fat and saturated fat. Fiber and sugar are also carbohydrates. Value of fruit and vegetables accounts for value over and beyond macronutrient characteristics of the fruit and vegetables.
## Table 6: Decomposition of Nutrition-Income Relationship by Household Demographics

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Unconditional Estimates</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(Household income)</td>
<td>0.127</td>
<td>0.135</td>
</tr>
<tr>
<td></td>
<td>(0.00478)***</td>
<td>(0.00861)***</td>
</tr>
<tr>
<td><strong>Panel B: Full Model</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(Household income)</td>
<td>0.0901</td>
<td>0.0930</td>
</tr>
<tr>
<td></td>
<td>(0.00597)***</td>
<td>(0.0104)***</td>
</tr>
<tr>
<td>ln(Years education)</td>
<td>0.578</td>
<td>0.609</td>
</tr>
<tr>
<td></td>
<td>(0.0288)***</td>
<td>(0.0500)***</td>
</tr>
<tr>
<td>Age indicators</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-0.18</td>
<td>-0.11</td>
<td></td>
</tr>
<tr>
<td>1(Have children)</td>
<td>0.0438</td>
<td>0.0290</td>
</tr>
<tr>
<td></td>
<td>(0.0102)***</td>
<td>(0.0180)</td>
</tr>
<tr>
<td>1(White)</td>
<td>0.114</td>
<td>0.0445</td>
</tr>
<tr>
<td></td>
<td>(0.0136)***</td>
<td>(0.0261)***</td>
</tr>
<tr>
<td>-0.003</td>
<td>-0.001</td>
<td></td>
</tr>
<tr>
<td>1(Black)</td>
<td>-0.239</td>
<td>-0.333</td>
</tr>
<tr>
<td></td>
<td>(0.0173)***</td>
<td>(0.0324)***</td>
</tr>
<tr>
<td>0.034</td>
<td>0.040</td>
<td></td>
</tr>
<tr>
<td>1(Married)</td>
<td>0.0521</td>
<td>0.0521</td>
</tr>
<tr>
<td></td>
<td>(0.00870)***</td>
<td>(0.0154)***</td>
</tr>
<tr>
<td>Employed</td>
<td>-0.116</td>
<td>-0.0587</td>
</tr>
<tr>
<td></td>
<td>(0.105)</td>
<td>(0.167)</td>
</tr>
<tr>
<td>-0.19</td>
<td>-0.08</td>
<td></td>
</tr>
<tr>
<td>Weekly work hours</td>
<td>0.00589</td>
<td>0.00432</td>
</tr>
<tr>
<td></td>
<td>(0.00311)*</td>
<td>(0.00499)</td>
</tr>
<tr>
<td>0.34</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Annual household calorie need (000s)</td>
<td>-0.0124</td>
<td>-0.0149</td>
</tr>
<tr>
<td></td>
<td>(0.00168)***</td>
<td>(0.00307)***</td>
</tr>
<tr>
<td>-0.09</td>
<td>-0.09</td>
<td></td>
</tr>
<tr>
<td>Health occupation</td>
<td>-0.0506</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0318)</td>
<td></td>
</tr>
<tr>
<td>-0.0064</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food knowledge</td>
<td>0.0156</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00698)**</td>
<td></td>
</tr>
<tr>
<td>0.0103</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>542,216</td>
<td>213,784</td>
</tr>
</tbody>
</table>

Notes: These regressions use 2004-2015 Nielsen Homescan data at the household-by-year level. The dependent variable is Health Index, our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. All regressions also control for year indicators. Each variable’s contribution to the nutrition-income relationship, i.e. \( \delta_k \) from Equation (31), is in italics. Observations are weighted for national representativeness. Robust standard errors, clustered by household, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
Figures

Figure 1: 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 2006-2015 Nielsen RMS data. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. 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 2: Healthfulness of Grocery Purchases by Household Income

Notes: This figure presents Nielsen Homescan data for 2004-2015. The x-axis presents nominal income bins; household incomes larger than $100,000 are coded as $125,000. Sugary drinks is the share of calories from sugar-sweetened beverages, 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 Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Observations are weighted for national representativeness.
Figure 3: Trends in Healthfulness of Grocery Purchases by Household Income

Notes: This figure presents Nielsen Homescan data for 2004-2015. The x-axis presents nominal income bins; household incomes larger than $100,000 are coded as $125,000. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Observations are weighted for national representativeness.
Figure 4: Store Average Healthfulness and Size by Zip Code Median Income

Notes: This figure uses Nielsen RMS data for 2006-2015. We constructed the calorie-weighted share of UPCs that are sugar-sweetened beverages, the calorie-weighted share of bread, buns, and rolls UPCs that are whole grain, the calorie-weighted share of UPCs that are produce, and the calorie-weighted mean Health Index, across all UPCs offered in each store. The left four panels of this figure present the means of these variables across stores within categories of zip code median income. The right two panels present revenues and UPC counts for the mean store in each zip code income category.
Figure 5: Relative Prices of Healthy and Unhealthy UPCs by Zip Code Median Income

Notes: This figure shows the average price per calorie across 2012 RMS transactions by income bin; household incomes larger than $100,000 are coded as $125,000. “Produce” includes packaged fresh, canned, and dried produce. The income group cutoffs are 25, 50, and 75 thousand dollars.
Notes: This figure presents the $\tau_{[0,10)^q}$ parameters and 95 percent confidence intervals from estimates of Equation (3): the effects of entry by several large supermarket chains, using 2004-2015 household-by-quarter Homescan data. All regressions control for household demographics (natural log of income, natural log of years of education, age indicators, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need), Census division-by-quarter of sample indicators, and household-by-Census tract fixed effects. The top two panels present effects on expenditure shares, in units of percentage points. The bottom two panels present effects on Health Index, our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. The dashed vertical line is the last quarter before entry. Observations are not weighted for national representativeness.
Figure 7: **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 household income less than $25,000. “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 8: Event Study: Health Index Changes in Mover Households

Notes: Using 2004-2015 Homescan data, these figures present the $\tau_y$ parameters and 95 percent confidence intervals from estimates of Equation (5): associations between household-level Health Index and the difference in average local Health Index between post-move and pre-move locations. All regressions control for household demographics (natural log of income, natural log of years of education, age indicators, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need), year indicators, and household fixed effects. Each figure superimposes three different estimates identified off of balanced panels for different windows around the move. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Observations are not weighted for national representativeness.
Figure 9: Predicted Health Index for Each Income Group

Notes: Each category on the x-axis represented a separate counterfactual calculation. The base category measures the health index for each income group when each group retains their true preferences and face their own local supply conditions. The second category sets all prices to those observed in income group 4. The third category sets all prices and product nutrient characteristics to those in group 4. The fourth and fifth categories add on the nutrient preferences and product group, respectively.
Figure 10: Percent of Health Index Differences From Supply & Demand Factors

Notes: Units measure the percent of the health index difference between the poorest and the richest income group that is caused by each factor: product group prices, product group nutrients, product group preferences, and nutrient preferences.
Online Appendix: Not for Publication

The Geography of Poverty and Nutrition: Food Deserts and Food Choices Across the United States

_Hunt Allcott, Rebecca Diamond, and Jean-Pierre Dubé_
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 items with UPCs, which are the items that we observe outside the Magnet subsample. “Produce” includes fresh, dried, canned, and frozen produce. Observations are weighted for national representativeness.

A.B Health Index

Our raw Health Index $H(x)$ is the sum of good minus bad nutrients, weighting each by its recommended daily intake (RDI): $H(x) = \sum_k G_k \frac{g_k}{r_k} - (1 - G_k) \frac{g_k}{r_k}$, where $g_k$ is the grams of macronutrient $k$, $r_k$ is the RDI for a normal adult, and $G_k$ takes value 1 for “good” macronutrients and 0 for “bad” macronutrients. Appendix Table A1 presents the RDI $r_k$ used to construct the Health Index. For example, $H(x)$ would take value 1 for a UPC (say three cups of vegetables) that exactly satisfied the RDI of one “good” macronutrient, or -1 for a UPC that contained the maximum RDI of one “bad” macronutrient.

In the 2004-2015 Homescan data, the mean household-level raw Health Index per 1000 calories is -2.79, and the within-year standard deviation is 0.72. The Health Index reported in the paper is re-normalized to mean 0, standard deviation one across households.
## Table A1: Health Index Function

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Recommendation</th>
<th>Intake (grams)</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fruits and vegetables</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fruits</td>
<td>Increase</td>
<td>320</td>
<td>Two cups/day (Food Patterns); 160 g/cup</td>
</tr>
<tr>
<td>Vegetables</td>
<td>Increase</td>
<td>390</td>
<td>Three cups/day (Food Patterns); 130 g/cup</td>
</tr>
<tr>
<td><strong>All other items</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>Increase</td>
<td>51</td>
<td>51 grams/day (DRI)</td>
</tr>
<tr>
<td>Fiber</td>
<td>Increase</td>
<td>29.5</td>
<td>29.5 grams/day (DRI)</td>
</tr>
<tr>
<td>Sugar</td>
<td>Reduce</td>
<td>32.8</td>
<td>45% of 282 calories/day from sugar+sat. fat (Food Patterns)</td>
</tr>
<tr>
<td>Saturated fat</td>
<td>Reduce</td>
<td>17.2</td>
<td>55% of 282 calories/day from sugar+sat. fat (Food Patterns)</td>
</tr>
<tr>
<td>Sodium</td>
<td>Reduce</td>
<td>2.3</td>
<td>2300 mg/day (Dietary Guidelines)</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>Reduce</td>
<td>0.3</td>
<td>300 mg/day (Dietary Guidelines)</td>
</tr>
</tbody>
</table>

Notes: This table presents the Recommended Daily Intakes (RDI) for each attribute. The Health Index \( H(x) = \sum_k G_k \frac{g_k}{r_k} - (1 - G_k) \frac{g_k}{r_k} \), where \( g_k \) is the grams of macronutrient \( k \), \( r_k \) is the RDI for a normal adult, and \( G_k \) takes value 1 for macronutrients recommended to “Increase” and 0 for macronutrients recommended to “Reduce.”

## Table A2: Correlations Between Health Index and Its Components in Homescan

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Correlation with Health Index</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fruits and vegetables</strong></td>
<td></td>
</tr>
<tr>
<td>Fruits</td>
<td>0.33</td>
</tr>
<tr>
<td>Vegetables</td>
<td>0.34</td>
</tr>
<tr>
<td><strong>All other items</strong></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>0.52</td>
</tr>
<tr>
<td>Fiber</td>
<td>0.64</td>
</tr>
<tr>
<td>Sugar</td>
<td>-0.75</td>
</tr>
<tr>
<td>Saturated fat</td>
<td>-0.05</td>
</tr>
<tr>
<td>Sodium</td>
<td>-0.19</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>-0.008</td>
</tr>
</tbody>
</table>

Notes: Using Homescan household-by-year data for 2004-2015, this table presents the correlation coefficients between Health Index and its components, using data in units of grams per 1000 calories consumed. Observations are weighted for national representativeness.
Table A3: Pooled OLS vs. Within-Household Income Variation

<table>
<thead>
<tr>
<th>Sample:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(Household income)</td>
<td>0.130</td>
<td>0.0198</td>
<td>0.0160</td>
<td>0.0134</td>
</tr>
<tr>
<td>(0.00506)***</td>
<td>(0.00484)***</td>
<td>(0.00492)***</td>
<td>(0.00580)**</td>
<td></td>
</tr>
<tr>
<td>Household-by-Census tract fixed effects</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Household demographics</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>603,230</td>
<td>603,230</td>
<td>603,230</td>
<td>516,170</td>
</tr>
<tr>
<td>Income coefficient/column 1 coefficient</td>
<td>1</td>
<td>0.15</td>
<td>0.12</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Notes: This table presents regressions of Health Index on natural log of household income and year indicators using Nielsen Homescan data for 2004-2015. Columns 2-4 also include household-by-Census tract fixed effects, and columns 3 and 4 also include household demographics (natural log of years of education, age indicators, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need). The sample is restricted to households observed in two or more years; column 4 additionally excludes observations with household income above $100,000. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Observations are weighted for national representativeness. Robust standard errors, clustered by household, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
B Appendix to Stylized Facts Section

B.A Additional Figures

Figure A2: Macronutrient Purchases by Household Income

Notes: Presents calorie-weighted average macronutrient contents of purchases using Nielsen Homescan data for 2004-2015. The x-axis presents nominal income bins; household incomes larger than $100,000 are coded as $125,000. Observations are weighted for national representativeness.
Figure A3: **Magnet Subsample: Healthful Purchases by Household Income**

Notes: Nielsen Homescan data, magnet subsample, for 2004-2006. The x-axis presents nominal income bins; household incomes larger than $100,000 are coded as $125,000. This parallels Figure 2, except using the magnet subsample which also records purchases of non-UPC items such as bulk produce. Produce is the share of calories from fresh, canned, dried, and frozen fruits and vegetables, and Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Observations are weighted for national representativeness.
Figure A4: Trends in Macronutrient Purchases by Household Income

Notes: Presents calorie-weighted average macronutrient contents of purchases using Nielsen Homescan data for 2004-2015. The x-axis presents nominal income bins; household incomes larger than $100,000 are coded as $125,000. On each plot, the three lines plot 2004-2007, 2008-2011, and 2012-2015 averages, respectively, in light, medium, and dark lines. Observations are weighted for national representativeness.
Figure A5: **Store Average Healthfulness by Zip Code Median Income**

Notes: Using Nielsen RMS data for year 2012, we constructed calorie-weighted mean macronutrient content across all UPCs offered in each store. This figure presents the means of these variables within categories of zip code median income. This parallels Figure 2 in the text.
Figure A6: 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 household income less than $25,000. “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.
B.B Low-income neighborhoods have relatively more unhealthful store types

Using the Zip Code Business Patterns data for 2004-2015, Appendix Figure A7 plots the average count of stores by channel type for zip codes by median income category. Zip codes vary substantially in area and population, so this figure normalizes store counts per 10,000 residents; the mean zip code has 12,000 residents. Lower-income zip codes have more stores per capita of all channel types, with two exceptions. First, the concentration of large grocery stores per capita is sharply monotonically increasing in median income, consistent with Powell et al. (2007). Second, the concentration of supercenters and club stores takes an inverted-U shape, with many fewer per capita in the very lowest-income zip codes.

Figure A7: Store Counts by Zip Code Median Income

Notes: This figure presents mean store counts per 10,000 residents by zip code income category using data from Zip Code Business Patterns, averaged over 2004-2015. Large (small) grocers are defined as those with 50 or more (fewer than 50) employees.
C Appendix to Section IV

C.A Additional Figures and Tables for Entry Event Study

Figure A8: 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 (3): the effects of entry by several large supermarket chains, using 2004-2015 household-by-quarter Homescan data. All regressions control for household demographics (natural log of income, natural log of years of education, age indicators, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need), Census division-by-quarter of sample indicators, and household-by-Census tract fixed effects. The top two panels present effects on expenditure shares, in units of percentage points. The bottom two panels present effects on Health Index, our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. The dashed vertical line is the last quarter before entry. Observations are not weighted for national representativeness.
Table A4: **Effects of Supermarket Entry**

<table>
<thead>
<tr>
<th>Sample:</th>
<th>Full Sample</th>
<th>Income &lt; $25,000</th>
<th>“Food Desert” Zip Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expenditure shares at store type:</td>
<td>Conv/ Drug Stores</td>
<td>Other Mass Merchants</td>
<td>Conv/ Drug Stores</td>
</tr>
<tr>
<td>Post entry: 0-10 minutes</td>
<td>-0.0887 <strong>(0.0352)</strong></td>
<td>-0.515 <strong>(0.0641)</strong></td>
<td>-0.265 <strong>(0.125)</strong></td>
</tr>
<tr>
<td>Post entry: 10-15 minutes</td>
<td>-0.0540 <strong>(0.0254)</strong></td>
<td>-0.113 <strong>(0.0434)</strong></td>
<td>-0.120 <strong>(0.0904)</strong></td>
</tr>
<tr>
<td>Observations</td>
<td>2,606,413</td>
<td>2,606,413</td>
<td>400,690</td>
</tr>
<tr>
<td>Dependent var. mean</td>
<td>2.6</td>
<td>5.3</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Panel B: **Effects on Health Index Using Alternative Food Desert Definitions**

<table>
<thead>
<tr>
<th>Sample:</th>
<th>&lt; 1000 Produce UPCs</th>
<th>No Medium Groceries</th>
<th>Three-Mile Radius</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post entry: 0-10 minutes</td>
<td>-0.00977 <strong>(0.0182)</strong></td>
<td>-0.0131 <strong>(0.0192)</strong></td>
<td>0.00550 <strong>(0.0200)</strong></td>
</tr>
<tr>
<td>Post entry: 10-15 minutes</td>
<td>0.0336 <strong>(0.0109)</strong></td>
<td>0.0303 <strong>(0.0115)</strong></td>
<td>0.0462 <strong>(0.0133)</strong></td>
</tr>
<tr>
<td>Observations</td>
<td>399,291</td>
<td>371,862</td>
<td>475,597</td>
</tr>
</tbody>
</table>

Notes: This table uses 2004-2015 Nielsen Homescan data at the household-by-quarter level. The table parallels Table 3, except Panel A presents effects on expenditure shares at alternative channel types, and Panel B uses alternative definitions of a “food desert.” In Panel B, columns 1 and 2 limit the sample to zip codes with fewer than 1000 produce UPCs available in 2003, as predicted by applying RMS data from Table 2 to Zip Code Business Patterns data; columns 3 and 4 also exclude any zip codes with grocery stores employing between 10 and 49 employees in 2003; columns 5 and 6 define 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 in 2003. Expenditure shares are in units of percentage points. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Reported independent variables are indicators for whether a specific retailer has 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, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need), 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, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
Figure A9: **Channel Type Expenditure Shares by Household Income**

Notes: This figure uses Nielsen Homescan data for 2004-2015. The x-axis presents nominal income bins; household incomes larger than $100,000 are coded as $125,000. Another 5-6 percent of expenditures are at channels not plotted, including bakeries, butchers, candy stores, liquor stores, fruit stands, and fish markets; this proportion is fairly constant by income. Observations are weighted for national representativeness.
C.B Entry by All Retailers Using Zip Code Business Patterns

To complement the event study estimates in the body of the paper, we present alternative specifications that measure entry using the Zip Code Business Patterns (ZBP) data.

Panel A of Appendix Table A5 shows that the Zip Code Business Patterns data date openings of specific supercenters in the correct year 50 to 80 percent of the time, although they are sometimes recorded a year later and sometimes in a broader “general merchandise” NAICS code (452) instead of the specific “supercenter and club store” NAICS code (452910).

The entry event study regression is analogous to Equation (2). Define $S_{zt}$ and $G_{zt}$, respectively, as the count of supercenters/club stores and large (at least 50 employee) grocery stores in zip code $z$ in year $t$. Using household-by-year data and now denoting $\mu_{dt}$ as Census division-by-year indicators, the regression is:

$$Y_{it} = \tau S_{zt} + \tau G_{zt} + \beta X_{it} + \mu_{dt} + \phi_{ic} + \varepsilon_{it}$$ (32)

Standard errors are again clustered by household, and observations are again all weighted equally.

Appendix Table A6 presents results. The structure is similar to that of Table 3: Panel A presents effects on expenditure shares, while Panel B presents effects on healthful eating.

Columns 1-3 present estimates for the full sample. Columns 1 and 2 confirm that the ZBP data contain meaningful information. Column 1 shows that conditional on household fixed effects, a larger count of large grocery stores and/or a smaller count of supercenters and club stores in the zip code are both strongly positively associated with higher expenditure share at chain groceries. Column 2 shows the opposite: fewer grocery stores and more supercenters are strongly positively associated with higher expenditures at supercenters and club stores. Column 3 presents effects on combined expenditure shares for all grocery stores, supercenters, and club stores. Columns 4-6 and 7-9 present estimates for the low-income and food desert subsamples. As in Table 3, effects of entry on expenditures generally larger in food deserts. Also as in Table 3, we see that entry by a large grocery retailer substantially diverts sales from other supermarkets, so the effects on combined expenditures at grocery stores, supercenters, and club stores are limited. Appendix Table A7 shows that most of this diversion is from other mass merchants; there is no statistically significant diversion from drug and convenience stores.

The bottom panel shows no statistically significant effect of the number of large grocers and supercenters/clubs on Health Index. With 95 percent confidence, we can bound the effects on low-income households’ Health Index at less than 0.011 standard deviations per large grocery store and 0.052 standard deviations per supercenter or club store. Appendix Table A7 shows that under all alternative definitions of “food deserts,” the number of local large grocers, supermarkets, and club stores has no statistically or economically significant effect on Health Index.

One reason to prefer the earlier regressions with specific known retailers is that we have high
confidence that entry dates are correctly measured. We can also imagine using the true supercenter entry dates as an instrument for ZBP data, which are measured with error. Panel B of Appendix Table A5 shows that the “first stages” of such a regression have coefficients around 0.9 and 0.66 for two different supercenter chains. If the average retailer in ZBP is measured with equal or perhaps somewhat more error than the less well-measured supercenter chain, this suggests that our bounds in the paragraph above should be increased by 50 to 100 percent due to measurement error. Even after this adjustment, however, our results in Tables 3 and A6 suggest that having a supermarket nearby explains at most only a small share of the differences in nutritional decisions between low- and high-income households.
Table A5: **Zip Code Business Patterns Accuracy Check with Known Entry Dates**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable</td>
<td>General</td>
<td></td>
</tr>
<tr>
<td>count of channel type</td>
<td>Supercenter</td>
<td>Merchandise</td>
</tr>
<tr>
<td><strong>Panel A: Difference Estimator</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supercenter Chain 1:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-year lead</td>
<td>0.00478</td>
<td>0.00629</td>
</tr>
<tr>
<td></td>
<td>(0.00622)</td>
<td>(0.0166)</td>
</tr>
<tr>
<td>1-year lead</td>
<td>0.0375</td>
<td>0.0577</td>
</tr>
<tr>
<td></td>
<td>(0.00978)**</td>
<td>(0.0164)***</td>
</tr>
<tr>
<td>Entry year</td>
<td>0.562</td>
<td>0.821</td>
</tr>
<tr>
<td></td>
<td>(0.0199)***</td>
<td>(0.0210)***</td>
</tr>
<tr>
<td>1-year lag</td>
<td>0.208</td>
<td>0.0821</td>
</tr>
<tr>
<td></td>
<td>(0.0169)***</td>
<td>(0.0191)***</td>
</tr>
<tr>
<td>2-year lag</td>
<td>0.0777</td>
<td>0.0133</td>
</tr>
<tr>
<td></td>
<td>(0.0127)***</td>
<td>(0.0145)</td>
</tr>
<tr>
<td>Supercenter Chain 2:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-year lead</td>
<td>0.0158</td>
<td>-0.0172</td>
</tr>
<tr>
<td></td>
<td>(0.0298)</td>
<td>(0.0421)</td>
</tr>
<tr>
<td>1-year lead</td>
<td>0.0133</td>
<td>-0.0451</td>
</tr>
<tr>
<td></td>
<td>(0.0222)</td>
<td>(0.0462)</td>
</tr>
<tr>
<td>Entry year</td>
<td>0.0621</td>
<td>0.480</td>
</tr>
<tr>
<td></td>
<td>(0.0327)*</td>
<td>(0.0623)***</td>
</tr>
<tr>
<td>1-year lag</td>
<td>0.172</td>
<td>0.0701</td>
</tr>
<tr>
<td></td>
<td>(0.0413)***</td>
<td>(0.0594)</td>
</tr>
<tr>
<td>2-year lag</td>
<td>0.133</td>
<td>0.0918</td>
</tr>
<tr>
<td></td>
<td>(0.0514)***</td>
<td>(0.0599)</td>
</tr>
<tr>
<td>Observations</td>
<td>264,734</td>
<td>264,734</td>
</tr>
<tr>
<td><strong>Panel B: Fixed Effects Estimator</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post entry: chain 1</td>
<td>0.902</td>
<td>0.932</td>
</tr>
<tr>
<td></td>
<td>(0.0138)***</td>
<td>(0.0227)***</td>
</tr>
<tr>
<td>Post entry: chain 2</td>
<td>0.667</td>
<td>0.665</td>
</tr>
<tr>
<td></td>
<td>(0.0365)***</td>
<td>(0.0659)***</td>
</tr>
<tr>
<td>Observations</td>
<td>297,966</td>
<td>297,966</td>
</tr>
</tbody>
</table>

Notes: Data are at the zip code-by-year level. All regressions include year indicators; fixed effects regressions have zip code fixed effects. Robust standard errors, clustered by zip code, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
Table A6: Effects of Supermarket Entry Using Zip Code Business Patterns

**Panel A: Effects on Expenditure Shares**

<table>
<thead>
<tr>
<th>Sample:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expend. shares at store type:</td>
<td>Full Sample</td>
<td>Income &lt; $25,000</td>
<td>“Food Desert” Zip Codes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large grocers</td>
<td>Chain 0.329 (0.0718)***</td>
<td>Super/Club -0.382 (0.0615)***</td>
<td>Grocery/Club 0.0259 (0.0414)</td>
<td>Chain 0.256 (0.207)</td>
<td>Super/Club 0.430 (0.173)***</td>
<td>Grocery/Club -0.0772 (0.136)</td>
<td>Chain 0.522 (0.299)*</td>
<td>Super/Club 0.705 (0.242)***</td>
<td>Grocery/Club 0.129 (0.167)</td>
</tr>
<tr>
<td>Supers/clubs</td>
<td>Chain -1.896 (0.172)***</td>
<td>Super/Club 2.903 (0.160)***</td>
<td>Grocery/Club 0.692 (0.0939)***</td>
<td>Chain 3.553 (0.480)***</td>
<td>Super/Club 1.057 (0.451)***</td>
<td>Grocery/Club -2.472 (0.311)***</td>
<td>Chain 3.452 (0.591)***</td>
<td>Super/Club 0.629 (0.570)***</td>
<td>Grocery/Club (0.291)**</td>
</tr>
<tr>
<td>Observations</td>
<td>664,302</td>
<td>102,462</td>
<td>163,747</td>
<td>102,462</td>
<td>163,747</td>
<td>163,747</td>
<td>163,747</td>
<td>163,747</td>
<td></td>
</tr>
<tr>
<td>Dep. var. mean</td>
<td>58</td>
<td>26</td>
<td>88</td>
<td>56</td>
<td>23</td>
<td>86</td>
<td>53</td>
<td>28</td>
<td>88</td>
</tr>
</tbody>
</table>

**Panel B: Effects on Health Index**

<table>
<thead>
<tr>
<th>Sample:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample:</td>
<td>Full Sample</td>
<td>Income &lt; $25,000</td>
<td>“Food Desert” Zip Codes</td>
</tr>
<tr>
<td>Large grocers</td>
<td>0.00102 (0.00254)</td>
<td>-0.00303 (0.00738)</td>
<td>-0.00518 (0.0104)</td>
</tr>
<tr>
<td>Supers/clubs</td>
<td>0.00667 (0.00551)</td>
<td>0.0217 (0.0155)</td>
<td>0.0184 (0.0168)</td>
</tr>
<tr>
<td>Observations</td>
<td>664,302</td>
<td>102,462</td>
<td>163,747</td>
</tr>
</tbody>
</table>

Notes: This table uses 2004-2015 Nielsen Homescan data at the household-by-year level. “Food Desert” zip codes are those with no grocery stores with 50 or more employees, supercenters, or club stores in 2003. Expenditure shares are in units of percentage points. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Reported independent variables are the count of stores by channel type in the household’s zip code. All regressions control for household demographics (natural log of income, natural log of years of education, age indicators, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need), Census division-by-year indicators, and household-by-Census tract fixed effects. Observations are not weighted for national representativeness. Robust standard errors, clustered by household, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
Table A7: Effects of Supermarket Entry Using Zip Code Business Patterns

Panel A: Effects on Expenditure Shares at Other Store Types

<table>
<thead>
<tr>
<th>Sample:</th>
<th>Full Sample</th>
<th>Income &lt; $25,000</th>
<th>“Food Desert” Zip Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expenditure shares at store type:</td>
<td>Conv./ Other Mass</td>
<td>Conv./ Other Mass</td>
<td>Conv./ Other Mass</td>
</tr>
<tr>
<td>Large grocers</td>
<td>-0.0145</td>
<td>0.0737</td>
<td>-0.0352</td>
</tr>
<tr>
<td></td>
<td>(0.0173)</td>
<td>(0.0278)***</td>
<td>(0.0613)</td>
</tr>
<tr>
<td>Supercenters/clubs</td>
<td>-0.0216</td>
<td>-0.717</td>
<td>-0.172</td>
</tr>
<tr>
<td></td>
<td>(0.0348)</td>
<td>(0.0641)***</td>
<td>(0.130)</td>
</tr>
<tr>
<td>Observations</td>
<td>664,302</td>
<td>664,302</td>
<td>102,462</td>
</tr>
<tr>
<td>Dependent var. mean</td>
<td>2.6</td>
<td>5.2</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Panel B: Effects on Health Index Using Alternative Food Desert Definitions

<table>
<thead>
<tr>
<th>Sample:</th>
<th>&lt; 1000 Produce UPCs</th>
<th>No Medium Groceries</th>
<th>Three-Mile Radius</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large grocers</td>
<td>-0.00617</td>
<td>-0.00660</td>
<td>-0.0144</td>
</tr>
<tr>
<td></td>
<td>(0.0148)</td>
<td>(0.0161)</td>
<td>(0.0129)</td>
</tr>
<tr>
<td>Supers/clubs</td>
<td>0.0158</td>
<td>0.0133</td>
<td>0.0192</td>
</tr>
<tr>
<td></td>
<td>(0.0200)</td>
<td>(0.0204)</td>
<td>(0.0211)</td>
</tr>
<tr>
<td>Observations</td>
<td>104,451</td>
<td>98,256</td>
<td>125,399</td>
</tr>
</tbody>
</table>

Notes: This table uses 2004-2015 Nielsen Homescan data at the household-by-year level. The table parallels Table A6, except Panel A presents effects on expenditure shares at alternative channel types, and Panel B uses alternative definitions of a “food desert.” In Panel B, columns 1 and 2 limit the sample to zip codes with fewer than 1000 produce UPCs available in 2003, as predicted by applying RMS data from Table 2 to Zip Code Business Patterns data; columns 3 and 4 also exclude any zip codes with grocery stores employing between 10 and 49 employees in 2003; columns 5 and 6 define 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 in 2003. Expenditure shares are in units of percentage points. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Reported independent variables are the count of stores by channel type in the household’s zip code. All regressions control for household demographics (natural log of income, natural log of years of education, age indicators, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need), Census division-by-year indicators, and household-by-Census tract fixed effects. Observations are not weighted for national representativeness. Observations are not weighted for national representativeness. Robust standard errors, clustered by household, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
C.C Appendix to Movers Analysis

Figure A10: Event Study: Income Changes in Mover Households

(a) Moves Across Counties

(b) Moves Across Zip Codes

Notes: Using 2004-2015 Homescan data, these figures present the $\tau_y$ parameters and 95 percent confidence intervals from estimates of Equation (5): associations between natural log of household income and the difference in 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. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Observations are not weighted for national representativeness. The regressions are the same as in Figure 8, except with natural log of household income as the dependent variable and no controls for household demographics.
Figure A11: Event Study: Health Index Changes, without Demographic Controls

Notes: Using 2004-2015 Homescan data, these figures present the $\tau_y$ parameters and 95 percent confidence intervals from estimates of Equation (5): associations between household-level Health Index and the difference in 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. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Observations are not weighted for national representativeness. The regressions are the same as in Figure 8, except with no controls for household demographics.
Table A8: Association of Income with Local Area Health Index Using Movers

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zip code average Health Index</td>
<td>-0.00454</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00748)</td>
<td></td>
</tr>
<tr>
<td>County average Health Index</td>
<td>0.0608</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0212)***</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>655,977</td>
<td>664,156</td>
</tr>
</tbody>
</table>

Notes: This table uses 2004-2015 Nielsen Homescan data at the household-by-year level. The dependent variable is the natural log of household income. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation 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, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.

Table A9: Association of Coke Market Share with Local Area Coke Market Share Using Movers

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>County average Coke market share</td>
<td>0.123</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>(0.0379)***</td>
<td>(0.0378)***</td>
</tr>
<tr>
<td>Household demographics</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>383,555</td>
<td>383,555</td>
</tr>
</tbody>
</table>

Notes: This table uses 2004-2015 Nielsen Homescan data at the household-by-year level. 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, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need. All regressions also control for year indicators and household fixed effects. Observations are not weighted for national representativeness. Robust standard errors, clustered by household, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
D GMM Estimation

We can simplify the estimation problem by solving for our linear coefficients, $(\delta, \delta_{zip})$, as analytic functions of $\tilde{\beta}$:

$$ (\delta, \delta_{zip}) = P_D \left( \log(Y) - F \left( \tilde{p}, \tilde{a}; \tilde{\beta} \right) \right) $$

where $D$ is a matrix with the two dummy variables, $D_{hjT}$ and $D_{zip(h)}$, and $Z$ is a matrix with all of our instruments, $D$, $\tilde{a}_{hjTc}$ and $\Delta \tilde{p}_{j,T,m(h)}$, and $P_D = (D'ZWZ'D)^{-1} D'ZWZ'$. Substituting 33 back into 21, we can re-write the GMM estimator in terms of $\tilde{\beta}$:

$$ \hat{\tilde{\beta}} = \arg\min_{\tilde{\beta}} \left( \frac{1}{JHT} \sum_h \sum_T \sum_j g_{hjT} \left( \tilde{\beta}_h \right) \right)' W \left( \frac{1}{JHT} \sum_h \sum_T \sum_j g_{hjT} \left( \tilde{\beta}_h \right) \right). $$

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

$$ -2G \left( \tilde{\beta} \right)' WG \left( \tilde{\beta} \right) = 0 $$

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

$$ G \left( \tilde{\beta} \right) = \frac{1}{JHT} \begin{bmatrix} \tilde{a}' (I - DP_D) \\ \Delta \tilde{p}' (I - D_{zip}P_{D_{zip}}) \\ D' (I - DP_D) \end{bmatrix} \nabla_{\beta} F \left( \tilde{p}, \tilde{a}; \tilde{\beta} \right). $$

The covariance matrix of our full GMM estimator, $\Theta^{GMM} \equiv \left( \hat{\delta}, \hat{\delta}_{zip}, \hat{\tilde{\beta}} \right)$, is $\text{cov}(\Theta^{GMM}) = (G'WG)^{-1} G'W \Omega WG (G'WG)^{-1}$ with Jacobian matrix

$$ G = \frac{1}{JHT} \sum_h \sum_T \sum_j \begin{bmatrix} \tilde{\beta}'_j & -\tilde{a}_{hjT}D'_{zip(h)} & -\tilde{a}_{hjT} \nabla_{\beta} F \left( \tilde{p}_{hjT}, \tilde{a}_{hTjc}; \tilde{\beta}_h \right) \\ -\Delta \tilde{p}_{j,T,m(h)}D'_{hjT} & -\Delta \tilde{p}_{j,T,m(h)}D'_{zip(h)} & -\Delta \tilde{p}_{j,T,m(h)} \nabla_{\beta} F \left( \tilde{p}_{hjT}, \tilde{a}_{hTjc}; \tilde{\beta}_h \right) \\ -D_{hjT}D'_{hjT} & -D_{hjT}D'_{zip(h)} & -D_{hjT} \nabla_{\beta} F \left( \tilde{p}_{hjT}, \tilde{a}_{hTjc}; \tilde{\beta}_h \right) \\ -D_{zip(h)}D'_{hjT} & -D_{zip(h)}D'_{zip(h)} & -D_{zip(h)} \nabla_{\beta} F \left( \tilde{p}_{hjT}, \tilde{a}_{hTjc}; \tilde{\beta}_h \right) \end{bmatrix}.$$
and covariance matrix

\[ \Omega = E \left( g_{hjT}(\Theta) g_{hjT}(\Theta)' \right). \]

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

\[ \hat{\Omega} = \frac{1}{JHT} \sum_h \sum_{T,T'} \sum_{j,j'} g_{hjT}(\hat{\beta}) g_{hj'T'}(\hat{\beta})'. \]

### Appendix to Demand Model Estimates

#### Table A10: Calories-Price Elasticities (2SLS)

<table>
<thead>
<tr>
<th>Inc (\leq 25K)</th>
<th>(25K \leq Inc \leq 50K)</th>
<th>(50K \leq Inc \leq 75K)</th>
<th>(75K \leq Inc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\ln(\text{price-per-calorie}))</td>
<td>-0.966***</td>
<td>-1.010***</td>
<td>-1.042***</td>
</tr>
<tr>
<td></td>
<td>(0.0467)</td>
<td>(0.0393)</td>
<td>(0.0401)</td>
</tr>
<tr>
<td>F-stat (KP)</td>
<td>83.32</td>
<td>132.04</td>
<td>149.15</td>
</tr>
</tbody>
</table>

Notes: Standard errors are clustered by zip3-product group. Dependent variable is \(\ln(\text{calories})\) purchased for each household-product group-year. Magnitudes represent elasticity of calorie demand for a product group with respect to price per calorie. Ln prices are instrumented with our supplier and distribution cost instrument. See text for more details.
Figure A12: **Binscatter Plot of First Stage Price Regression**

Notes: Data is binned into 20 equal quartile bins and plotted. All income groups are included in this regression. Fixed effects for zip3 and product group are included and residualized out before plotting the regression.
F Persistence in Tastes and Early-Life Environment

One hypothesis that could explain why high- and low-income households eat differently is that eating habits developed early in life have lasting effects on preferences, and low-income people may grow up in less healthy eating environments. Identifying the effects of early-life environments is difficult, however: we need variation in early life environments that is exogenous to factors that would currently affect consumption. In this section, we present results of regressions that document a strong association between household Health Index and the average Health Index of purchases in the primary shopper’s birth state. We then present a speculative calculation that benchmarks the possible effect of early life environments on high- vs. low-income households’ healthy purchasing decisions under the (strong) assumption that this association is causal.

For this section, \( B_i \) is an indicator for whether the primary shopper in household \( i \) reports being born in a different state than the current state of residence in the survey implemented by Bronnenberg et al. (2012). \( H_n \) is again the average Health Index of groceries purchased in the household’s current county of residence, \( H_b \) is the average Health Index of groceries purchased in the primary shopper’s birth state, and \( \Delta_{bn} \equiv H_b - H_n \), the birth state minus current county Health Index. As we showed in Figure 1, there is significant geographic variation in Health Index of purchases across the U.S. Appendix Figure A13 shows the state level averages. States in the deep South tend to have lower Health Index, while the west coast and the Northeast have the highest Health Index, although there is some within-region variation as well. \( X_{it} \) are household demographic controls, \( \phi_g \) are geographic indicators at the level of the county or zip code, and \( \mu_t \) are year indicators. We estimate the following equation:

\[
Y_{it} = \tau \Delta_{bn} B_i + \alpha B_i + \beta X_{it} + \phi_g + \mu_t + \varepsilon_{it}.
\] (34)

Standard errors are again clustered by household, and observations are weighted equally, as this increases precision and movers are not nationally representative.

Table A11 presents estimates of Equation (34). Column 1 presents estimates excluding demographics \( X_{it} \), while column 2 adds \( X_{it} \), and column 3 switches \( \phi_g \) from county to zip code indicators. The estimates of \( \hat{\tau} \approx 0.175 \) imply that if we consider two households in the same location \( g \) with same demographics \( X_{it} \) whose primary shoppers were born in states where the Health Index differs by \( x \), the households now purchase groceries with Health Index that differs by 0.175\( x \). Columns 4 and 5 present additional regressions that measure whether this association is different for primary shoppers who were older when they moved away from their birth state or have lived in their current state for a longer time. Both interactions are statistically insignificant, but the standard errors are economically wide: we reject that \( \tau \) varies by more than 0.01 per year with 95 percent confidence. Thus, a household with primary shopper who moved at age 35 could have \( \tau \) about 0.3 larger than a household with primary shopper who moved at age five; this difference is
considerably larger than the average $\hat{\tau} \approx 0.175$ in column 3.\footnote{Appendix Figure A14 and Table A12 again extend the comparison to brand preferences by considering the case study of Coke’s share of Coke plus Pepsi purchases. Using Equation (34), the Coke market share $\hat{\tau} \approx 0.25$, and “exposure effects” are precisely estimated. One might expect that, when interpreting these as the causal effects of early life preferences, omitted variables could generate larger bias for healthy eating than for brand preferences. Thus, the fact that $\hat{\tau}$ is still larger for Coke than for Health Index implies that any causal effects of early-life environment are more important for soda brands than for healthy eating.}

$\tau$ reflects not just the causal effect of early-life environments, but also potential omitted variables. The fact that controlling for household demographics has little impact on $\hat{\tau}$ implies that any such other factors are not associated with observables. However, household demographics predict only a small amount of the variation in Health Index, so this leaves open the possibility for important confounders. We expect that any remaining selection on unobservables would make estimates of $\tau$ more positive than the true effects of birth state nutritional environment, so $\hat{\tau} \approx 0.175$ should be considered an upper bound on the causal effects of early-life environment.

Under this direction-of-bias assumption, we can calculate a bound on extent to which early-life environments contribute to the difference in Health Index between high- and low-income households. Given the tightly estimated zero effects of neighborhood environments from Section IV, for this calculation we consider the early life environment to be determined by the Health Index in the home, not in the local area. Using data from [tk cite], the average person with household income greater than $70,000 grew up in a household with income of $50,000, while the average person with household income less than $25,000 grew up in a household with income of $35,000. In the HMS data, the average child in a household with income of $50,000 has household normalized Health Index that is 0.10 larger than the average child in a household with income of $35,000. Thus, early life habits explain approximately $0.175 \times 0.1 = 0.0175 \approx 6\%$ of the high- vs. low-income difference in Health Index, and this is likely an upper bound due to the possibility of omitted variables that bias $\tau$ upwards.
Table A11: **Association Between Health Index and Health Index in Birth State**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Birth state - current county Health Index)</td>
<td>0.203</td>
<td>0.155</td>
<td>0.175</td>
<td>0.0917</td>
<td>0.188</td>
</tr>
<tr>
<td>×Mover</td>
<td>(0.0538)***</td>
<td>(0.0519)***</td>
<td>(0.0554)***</td>
<td>(0.0813)</td>
<td>(0.0913)**</td>
</tr>
<tr>
<td>(Birth state - current county Health Index)</td>
<td>0.00404</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>×Mover×Age of move</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00288)</td>
</tr>
<tr>
<td>(Birth state - current county Health Index)</td>
<td>-0.000491</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>×Mover×Years in current state</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00289)</td>
</tr>
<tr>
<td>Household demographics</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Geography indicators</td>
<td>County</td>
<td>County</td>
<td>Zip</td>
<td>Zip</td>
<td>Zip</td>
</tr>
<tr>
<td>Observations</td>
<td>231,377</td>
<td>231,377</td>
<td>231,377</td>
<td>231,377</td>
<td>231,377</td>
</tr>
</tbody>
</table>

Notes: This table uses 2004-2015 Nielsen Homescan data at the household-by-year level. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Household demographics are natural log of income, natural log of years of education, age indicators, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need. All regressions also control for a Mover indicator variable and year indicators. Observations are not weighted for national representativeness. Robust standard errors, clustered by household, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.
Figure A13: **Average Health Index of Store Purchases by State**

Notes: This figure presents the calorie-weighted average Health Index of purchases by state, using 2006-2015 RMS data. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. 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 A14: **Birth State Consumption Associations by Age of Move and Years Since Move**

Notes: These figures present estimates of Equation (34), allowing $\tau$ to vary by years since move and age of move, using Nielsen Homescan household-by-year data for 2004-2015. Coke market share equals Coke calories purchased / (Coke + Pepsi calories purchased); the sample mean is 0.56. Health Index is our overall measure of the healthfulness of grocery purchases, normalized to mean zero, standard deviation one across households. Observations are not weighted for national representativeness.
Table A12: Association Between Coke Purchase Share and Coke Market Share in Birth State

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Birth state - current county Coke share)</td>
<td>0.243</td>
<td>0.288</td>
<td>0.295</td>
<td>0.198</td>
<td>0.509</td>
</tr>
<tr>
<td>×Mover</td>
<td>(0.0327)***</td>
<td>(0.0319)***</td>
<td>(0.0337)***</td>
<td>(0.0481)***</td>
<td>(0.0557)***</td>
</tr>
<tr>
<td>(Birth state - current county Coke share)</td>
<td>0.00533</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>×Mover×Age of move</td>
<td>(0.00191)***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Birth state - current county Coke share)</td>
<td></td>
<td>-0.00798</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>×Mover×Years in current state</td>
<td></td>
<td></td>
<td>(0.00172)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household demographics</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Geography indicators</td>
<td>County</td>
<td>County</td>
<td>Zip</td>
<td>Zip</td>
<td>Zip</td>
</tr>
<tr>
<td>Observations</td>
<td>133,623</td>
<td>133,623</td>
<td>133,623</td>
<td>133,623</td>
<td>133,623</td>
</tr>
</tbody>
</table>

Notes: This table uses 2004-2015 Nielsen Homescan data at the household-by-year level. Coke market share equals Coke calories purchased / (Coke + Pepsi calories purchased); the sample mean is 0.56. Household demographics are natural log of income, natural log of years of education, age indicators, an indicator for whether the household includes children, race indicators, employment status, weekly work hours, and total calorie need. All regressions also control for a Mover indicator variable and year indicators. Observations are not weighted for national representativeness. Robust standard errors, clustered by household, in parentheses. *, **, ***: Statistically significant with 10, 5, and 1 percent confidence, respectively.