Doubled SNAP Dollars and Nudges: An Analysis of Two Pilot Programs Aimed at Increasing the Purchase of Healthy Foods

by

Danton Roberto Noriega-Goodwin

Department of Public Policy
Duke University

Date: ____________________________

Approved:

______________________________
Anna Gassman-Pines, Co-Supervisor

______________________________
Matthew Harding, Co-Supervisor

______________________________
Steven Sexton

______________________________
Kenneth Dodge

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Public Policy in the Graduate School of Duke University

2019
ABSTRACT

Doubled SNAP Dollars and Nudges: An Analysis of Two Pilot Programs Aimed at Increasing the Purchase of Healthy Foods

by

Danton Roberto Noriega-Goodwin

Department of Public Policy
Duke University

Date: __________________________

Approved:

______________________________
Anna Gassman-Pines, Co-Supervisor

______________________________
Matthew Harding, Co-Supervisor

______________________________
Steven Sexton

______________________________
Kenneth Dodge

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Public Policy in the Graduate School of Duke University

2019
Abstract

What people choose to eat is a public policy and health concern. Fresh produce and similarly healthy foods are often less preferred to unhealthy foods. Unhealthy foods can at times be cheaper and more accessible than healthy foods, making it difficult for consumers to avoid temptation at grocery or convenience stores. This dissertation is an analysis of two different pilot programs which aim to increase the purchase of healthy foods, specifically produce. The first pilot program is a financial incentive known as “Double Up Food Bucks”. The program is targeted towards SNAP participants, encouraging them to purchase more fresh produce by effectively doubling purchasing power. The second pilot program is a set of three behavioral nudges designed to increase the purchase of bananas in a convenience store environment. The impact of each pilot program was measured using a pre- and post-experiment difference-in-differences design. The results of both pilots are modest and support a growing body of evidence that traditional interventions, like financial incentives, and behavioral interventions, like nudges, can successfully increase healthy food purchases at the margin.
## Contents

List of Tables viii

List of Figures x

Acknowledgements xi

Introduction 1

1 An Evaluation of the Double Up Food Bucks Program 6

1.1 Introduction ...................................................... 6

1.1.1 Financial Incentives to Encourage Healthy Food Purchases 10

1.1.2 The Healthy Incentives Pilot .................................. 12

1.1.3 The Double Up Food Bucks Program .......................... 15

1.1.4 Double Up Food Bucks vs the Healthy Incentives Pilot ...... 17

1.1.5 Evaluating Double Up Food Bucks ............................ 18

1.1.6 Hypothesis ...................................................... 20

1.1.7 DUFB Mechanisms ............................................. 21

1.2 Data ............................................................... 28

1.2.1 Loyalty Card Concerns .......................................... 29

1.2.2 Gradual Decrease in SNAP Transactions ....................... 37
1.2.3 Overview of Store Selection and Expansion ................. 37
1.2.4 Selection and Expansion of DUFB Stores ................. 38
1.2.5 Expansion on Observables ................................................. 39
1.2.6 Selection of Control Stores ................................................. 43
1.2.7 An Additional 14 Stores ........................................... 45
1.2.8 Balance Between DUFB and non-DUFB Stores ................. 47
1.2.9 Loyalty Card Usage Rates ................................................. 48
1.2.10 Produce Spending ............................................................... 51
1.3 Empirical Strategy and Methods ........................................... 60
1.3.1 Store Switching Concerns ........................................... 63
1.3.2 Variables and Data Structure ........................................... 65
1.3.3 Methods ................................................................. 67
1.4 Results ................................................................. 70
1.4.1 Heterogeneous Effects ........................................... 75
1.5 Conclusion ................................................................. 80

2 Evaluating a USDA Pilot Program to Incentivize the Purchase of Fresh Produce Among SNAP Beneficiaries in Supermarkets 86
2.1 Introduction ................................................................. 86
2.2 Methods ................................................................. 88
2.3 Results ................................................................. 93
2.3.1 Fresh Produce ................................................................. 93
2.3.2 Sugar-Sweetened Beverages ........................................... 98
2.4 Discussion ................................................................. 99
# 3 Testing Behavioral Economic Nudges in a Chain of Convenience Store to Determine Impact on Healthy Food Choice Behavior

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>104</td>
</tr>
<tr>
<td>3.2 Methods</td>
<td>106</td>
</tr>
<tr>
<td>3.2.1 Study Design</td>
<td>106</td>
</tr>
<tr>
<td>3.2.2 Intervention</td>
<td>106</td>
</tr>
<tr>
<td>3.2.3 Ask Experiment</td>
<td>107</td>
</tr>
<tr>
<td>3.2.4 Pump Experiment</td>
<td>108</td>
</tr>
<tr>
<td>3.2.5 Bundling Experiment</td>
<td>109</td>
</tr>
<tr>
<td>3.3 Analysis Plan</td>
<td>111</td>
</tr>
<tr>
<td>3.3.1 Outcome measure</td>
<td>111</td>
</tr>
<tr>
<td>3.3.2 Baseline Data</td>
<td>112</td>
</tr>
<tr>
<td>3.3.3 Statistical Analysis</td>
<td>112</td>
</tr>
<tr>
<td>3.4 Results</td>
<td>116</td>
</tr>
<tr>
<td>3.4.1 Fidelity Checks</td>
<td>117</td>
</tr>
<tr>
<td>3.4.2 Change in Sales</td>
<td>117</td>
</tr>
<tr>
<td>3.5 Discussion</td>
<td>119</td>
</tr>
<tr>
<td>3.5.1 Limitations</td>
<td>120</td>
</tr>
<tr>
<td>3.5.2 Strengths</td>
<td>121</td>
</tr>
<tr>
<td>3.6 Recommendations</td>
<td>121</td>
</tr>
<tr>
<td>3.7 Conclusion</td>
<td>122</td>
</tr>
</tbody>
</table>

## Conclusion

## Appendix
List of Tables

1.1 Pooling Together all Transactions and Items Purchased. Grouped by Year. 30
1.2 Pooling Together all SNAP Transactions and Items Purchased. Grouped by Year. 31
1.3 Loyalty Transactions, SNAP vs Non-SNAP (Store-Month by Year). Percentages calculated out of all observations. 32
1.4 Loyalty Transactions, Pooled 32
1.5 Loyalty Transactions, SNAP Only 33
1.6 Year by Year DUFB Store Selection 40
1.7 Average Monthly Produce Spending per Loyalty ID 58
1.8 Average Monthly Non-Produce Spending per Loyalty ID 59
1.9 Average Monthly Proportion of Spending on Produce per Loyalty ID 60
1.10 Count of Loyalty IDs by Stores Visited in 2016 65
1.11 Sample of Rows from Data Set (Columns 1 - 11) 66
1.12 Sample of Rows from Data Set (Columns 12 - 15) 66
1.13 Produce Spending Triple Difference Estimates 71
1.14 Non-Produce Spending Triple Difference Estimates 71
1.15 Heterogeneous Produce Spending Triple Difference Estimates 79
2.1 Average Percentage of Total Dollars Spent on all Fresh Produce and Sugar-Sweetened Beverages Per Month 93
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2 Model Estimates for the Effect of the Double Up Food Bucks (DUFB) Program on the Average Percentage of Total Dollars Spent, 2015</td>
<td>95</td>
</tr>
<tr>
<td>2.3 Model Estimates for the Effect of the Double Up Food Bucks (DUFB) Program on the Average Percentage of Total Dollars Spent, 2016</td>
<td>96</td>
</tr>
<tr>
<td>3.1 Summary of Banana Sale Counts</td>
<td>115</td>
</tr>
<tr>
<td>3.2 Difference-in-Differences, Average Marginal Effects (2015 and 2016)</td>
<td>117</td>
</tr>
<tr>
<td>3.3 ANCOVA, Average Marginal Effects (2016)</td>
<td>117</td>
</tr>
<tr>
<td>3.4 CEM Match Matrix</td>
<td>129</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Example of Customer Receipt with DUFB Points Redeemed .............. 24
1.2 Loyalty Card Transactions as Percentage of all Transactions (Histogram) . 34
1.3 Loyalty Card Transactions as Percentage of all Transactions (Density) . 34
1.4 Loyalty Card Transactions as a Percentage of Total Monthly Transactions 35
1.5 Loyalty Card Transactions as Percentage of SNAP Transactions ........ 36
1.6 Month to Month US Household SNAP Enrollment, 2012 - 2017 ........ 38
1.7 Example Expansion Over Time from 2014 to 2016 (top to bottom) Using Fake Data .................................................. 42
1.8 Store Score vs DUFB Assignment ........................................ 46
1.9 Monthly Loyalty Card Transaction as a Percentage of Total Monthly Transactions per Store .............................................. 48
1.10 Average Number of Observed Loyalty Card Transaction per Customer Averaged per Store (Only SNAP Transactions) ............... 50
1.11 Average Number of Observed Loyalty Card Transaction per Customer Averaged per Store (non-SNAP Transactions) ............... 50
1.12 Average of Store Dollars Spent on Produce as a Percentage of All Dollars. 52
1.13 Average of Store SNAP Dollars Spent on Produce as a Percentage of SNAP Dollars. Monthly Observations from May 2014 - December 2016 Aggregated by Store and Grouped By DUFB Assignment ............... 53
1.14 Transformed Log-Normal Store Average of Monthly Customer Spending on Produce Grouped by DUFB Assignment, SNAP Status, and Month .... 57
Acknowledgements

I owe a great debt of gratitude to many people.

To my favorite statistics professor Kagba Suary, who celebrated loudest when I was accepted into PhD programs and will no doubt celebrate even louder when I finish. He has supported me since the beginning and I could always count on him to keep me motivated. To my advisor, Matthew Harding, who provided me with all the opportunities I needed to learn and grow as PhD student. He never stopped advocating for me and remains a fierce friend and ally, even after moving institutions, twice. To my unofficial mentor Randall Lewis, who when I most doubted the value of my intellectual interests provided me with much needed affirmation and valuable advice. To my aggressively supportive Sanford advisor, Anna Gassman-Pines, who stepped in at the most difficult time in my academic career and provided me a sense of belonging that rebuilt my confidence. To Lee Barnes, Oran Hesterman, and Alison Sutter for making my dissertation possible. Without them, I would have nothing to study. Finally, to the Sanford Department faculty and staff, for taking a chance on an unconventional graduate student.

To my parents, Eva and Cosme, who never quite understood why I decided to pursue a PhD, but were proud of me anyway. I tried my best to continue the family tradition of ending up ABD. The fact that you two didn’t care how things turned out always provided...
me with a healthy perspective; what mattered to you both is that I did whatever made me happiest.

Lastly, to Emily Rose, my lovely wife-to-be. No one has supported me more. No matter how my academic journey turned out, I would go back and do it all over again. Because the most important outcome of my graduate student career isn’t the PhD, it was meeting you.
Introduction

Individual dietary choices in the United States produce expensive negative externalities. Obesity, heart disease, type II diabetes, and other similar chronic conditions account for hundreds of thousands of deaths each year and cost the US health care system hundreds of billions of dollars annually (Cawley and Meyerhoefer, 2012). Poor diet is closely linked to these chronic conditions. Policy makers are therefore continuously searching for ways of improving the diet and lives of Americans at minimal economic cost.

Most public policy efforts to improve the American diet have historically fallen under two categories, guidelines and regulation (Welsh et al., 1992; Marc Law, 2004). It is normal for governments to recommend some form of a “healthy diet” to its citizens. In the United States, there was the Department of Agriculture’s famous “Food Pyramid” for two decades before being replaced by “MyPlate” in 2011 (Haven et al., 2006). But MyPlate, like all other guidelines before it, appears to have little sway on the dietary choices of most Americans.; most Americans, regardless of income, eat far too much meat and far too few fruits and vegetables (USDA, 2015). The other category, regulation, tends to focus on the health and safety of foods. Some regulation is uncontroversially restrictive, like if foods are dangerous to consumers because they contain bacteria or dangerous chemicals. But regulation falls short of banning the consumption of naturally occurring foods and ingredi-
ents, even if the long-term consequences of consuming certain foods and ingredients may negatively affect the health of most humans e.g. sugar and nicotine. Policy makers instead make the consumption of natural products known to cause long-term health problems more expensive. For example, alcohol and tobacco products are heavily taxed by most industrialized nations, both to discourage the consumption of such products and to also raise funds to pay for the adverse health outcomes caused by long-term consumption (Evans et al., 1999). Sugar is also having its moment as more is learned about its addictiveness and negative metabolic properties (Lustig et al., 2012; Zhen et al., 2011). Cities and countries—for example, Berkeley, CA and Mexico—have started taxing sugar-sweetened beverages as well as updating food labels with more obvious information about how much sugar should be consumed per day (Falbe et al., 2015; Colchero et al., 2016). Evidence suggests that taxes (financial disincentives) are effective methods of reducing the purchase, and consequently, consumption of almost any product (Novak and Brownell, 2011). And they are, like most taxes, very unpopular (Brownell, 2015).

Of course, the opposite—subsidies—are also effective but have not, until the last decade, been explored as a way for policy makers to incentivize the purchase of healthy foods. Policy makers have also started moving beyond classical economic theory and into the realm of behavioral economics to encourage the purchase and consumption of healthy foods. This dissertation researches and analyzes two recent efforts to increase the purchase of healthy foods without government regulation or guidelines. The first is a pilot program that follows a recent trend in the United States of targeted financial subsidies (financial
incentives) to certain consumers to encourage the purchase of healthier foods. Such financial incentives can be politically unpalatable but unpalatability decreases when financial subsidies are targeted towards more economically disadvantaged populations, like households that are eligible for Supplemental Nutrition Assistance Program (SNAP) (Blumenthal et al., 2014; Leung et al., 2013). The second is a pilot program that attempts to increase the purchase of healthy foods through behavioral nudges (Thaler and Sunstein, 2009). Both pilot programs are novel attempts by public policy researchers to encourage healthier food purchasing decisions.

In the first chapter of this dissertation I analyze the Double Up Food Bucks (DUFB) program, a financial incentive that doubles the fruit and vegetable purchasing power of SNAP households. My motivation for researching the impact of the DUFB program is straightforward. The purchase of fruits and vegetables acts a reasonable proxy for consumption. It is not a perfect proxy, but it is reasonable to assume that most a household cannot eat (consume) that which it does not purchase. This does not imply a household will eat (consume) all that purchases, but it is likely non-zero. Increasing produce spending can therefore be considered necessary for increasing the consumption of produce. If the DUFB program can notably increase fruit and vegetable spending, then it will have a non-negative impact on fruit and vegetable consumption.

Cost-effective interventions that increase fruit and vegetable consumption, particularly of low-income households like those participating in SNAP, is of continual interest for public health and policy makers. The DUFB program was considered a knock-out suc-
cess by policy makers after it was first launched in farmers markets in 2009, with SNAP benefits having been used more than 200,000 times by roughly 10,000 first time SNAP customers in 2013 alone (Network, 2014). However, less than 1% percent of SNAP benefits are used at farmer’s markets. To truly make an impact, the DUFB program had to prove it could be implemented at scale across supermarkets and grocery stores. The opportunity came when DUFB received a 5.17 million dollar FINI grant in 2014 to pilot an expansion into grocery stores. The pilot was launched into 2 stores in 2014, expanded into 5 stores in 2015, and then 17 stores in 2016. My involvement with the 2016 expansion provided a unique research opportunity to measure the impact, if any, that the DUFB incentive was having on the purchasing behavior of SNAP recipients. This paper moves beyond more traditional survey data research, using transaction-level data from a panel of individual shoppers. A triple-differences regression framework is used to measure the impact of the DUFB incentive on produce spending, exploiting the ability to link transactions to individuals and stores via loyalty ID numbers.

The second chapter of this dissertation is earlier work co-authored with Pasquale Rummo, Alex Parret, Matthew Harding, Oran Hesterman, and Brian Elbel. This second paper is included to supplement and support the findings of the first chapter. It is also included in the spirit of iteration and reproducibility (Duvendack et al., 2017; Goodman et al., 2016; Munafò et al., 2017). My co-authors and I explore the same question in the second chapter that I explore myself in the first, but using data from fewer stores. It also uses the same triple-difference regression framework to measure the effect of the DUFB incentive
but does not measure how different spending groups reacted to the DUFB incentive (aka “heterogeneous effects”). The major difference in the two papers is my involvement with the modeling part of the paper. I explicitly removed myself from helping with the modeling work in this second paper. I wanted another study to exist which used the data I worked so hard over many years to collect, clean, and curate, but where my own possible shortcomings and biases didn’t shape the modeling and findings. It was the only other way, given the restrictive data disclosure agreement, to have another study that used parts of the same dataset. In short, it was a way for me to pay homage to the reproducibility movement short of releasing a fully reproducible dissertation. Contributing to this paper was the first, and only, time I was able to observe other academics benefiting from the very data reproducibility and accessibility I championed.¹

In the third and final chapter, my co-authors—Molly De Marco, Lee Barnes, Madelaine Katz, Terry Hartman, and Matthew Harding—and I research whether convenience store customers can be influenced towards purchasing bananas over less healthy alternatives without financial incentives. We explore if customers can shift, on the margin, towards making healthier food choices nearest the moment impulsive snack-purchasing decisions are being made. The study adds to the literature of exploring non-intrusive, low-cost behavioral economics interventions (nudges) aimed at helping consumers make healthier purchasing decisions (List and Samek, 2015; Chance et al., 2014; Cohen et al., 2015). We research three different experiments randomly assigned to 32 stores and implemented in

¹Findings for both papers were finished around roughly the same time. Findings from one paper, therefore, did not, nor could not, influence the other.
different phases. The staggered implementations are used to construct three difference-in-difference models, one for each of the three experiments.
Chapter 1

An Evaluation of the Double Up Food Bucks Program

1.1 Introduction

Chronic conditions like obesity, heart disease, and other metabolic risk factors (stroke, type II diabetes, etc.) cost the US health care system between 200 to 400 billion dollars annually (Cawley and Meyerhoefer, 2012; Chatterjee et al., 2014). More importantly, these diseases account for hundreds of thousands of deaths each year. Heart disease alone is the leading cause of death for all persons in the US, with stroke fifth and diabetes seventh (National Center for Health Statistics, 2015). Diet is closely linked to these conditions, particularly obesity and cardiovascular disease. Strong evidence shows that a diet high in vegetables, fruits, nuts, unsaturated oils, fish, and poultry, but low in red and processed meat and sugar-sweetened foods and drinks, helps lower body weight, blood pressure, and the risk of cardiovascular disease (Mente et al., 2009; Nutrition Evidence Library, 2014). Improving the diet of Americans has therefore become an increasing priority for the United States, particularly for families in the Supplemental Nutrition Assistance (SNAP) program.

SNAP is a federal aid program administered by the Food and Nutrition Service (FNS), an agency of the U.S. Department of Agriculture (USDA). At 74 billion dollars in
FY2015 with 45.8 million participants, it is the largest food assistance program in the United States (USDA FNS, 2016b). To be eligible for SNAP, a household must be sufficiently budget constrained that hunger is considered likely without assistance. Eligibility is a function of countable resources, vehicle ownership and value, household size, gross or net monthly income, household composition, and meeting certain work requirements. Some eligibility requirements vary by state, but in general, a family with dependents, less than $2000 in countable resources, where the adults work at least part-time earning a gross (net) monthly income at or below 130% (100%) of the federal poverty line, is eligible to receive SNAP benefits. Aside from a few restrictions—no alcohol, tobacco, non-food items, ready-to-eat meals, or hot foods—households can use SNAP benefits to buy any foods for at-home consumption. Unfortunately, the purchasing patterns of the average SNAP household are not conducive to a healthy diet.

Research on the dietary patterns of households receiving SNAP benefits has found that they are significantly less likely to meet USDA dietary guidelines than the average US household and much more likely to consume unhealthy foods (Andreyeva et al., 2015; Nguyen et al., 2015; Wolfson and Bleich, 2015). A smaller set of research has found that SNAP households, at best, consume the same amount of unhealthy foods (e.g. sugar-sweetened beverages, baked goods, snacks, candy, etc) compared to SNAP-ineligible households (Todd and Ploeg, 2014; Hoynes et al., 2015). In other words, SNAP households consume foods that are less healthy or about the same as SNAP-ineligible households. This is a concerning result given that most US households, regardless of income, already purchase

\[1\] For more details, visit http://www.fns.usda.gov/snap/eligibility
and consume far too much meat and foods rich in sugars and fats, and far too few fruits, vegetables and whole grains (USDA, 2015; Frazão, 1999). The purpose of SNAP, however, is to keep struggling families from going hungry, not to ensure they consume the best possible diet. SNAP is designed to act like cash, helping families access more food than they could otherwise do so without assistance (Hoynes et al., 2015). It is therefore not a failing of the SNAP program if benefits are used to purchase unhealthy foods.

The SNAP program could be changed such that it could continue satisfying its role as an anti-hunger program while simultaneously encouraging healthier purchases. Blumenthal et al. (2014) and Leung et al. (2013) both surveyed a field of stakeholders and policy experts in the SNAP program about what they would do to improve the dietary quality of purchases. In both studies, restricting the purchase of unhealthy foods (e.g. sugar-sweetened beverages) and promoting healthy purchases through monetary incentives were the two most popular improvements (i.e. ranked the highest or most often suggested).2

One common suggestion is to restrict the SNAP program to the same set of eligible foods as the Special Supplemental Nutrition Program for Woman, Infants, and Children (WIC) (Dinour et al., 2007). The WIC program provides food vouchers that limit households to a select group of food products. These food products are specifically selected to ensure women and their children receive nutritious, healthy foods. In other words, the WIC program, by design, places restrictions on food choices by defining a list of eligible items, as opposed to the SNAP program, which defines a list of ineligible items. Another common,

---

2It should be mentioned that there were other, less popular recommendations, such as modifying how SNAP benefits are distributed and improving nutrition education. For more details, see Blumenthal et al. (2014) and Leung et al. (2013).
and simpler, suggestion is to expand the existing list of ineligible items (e.g. alcohol) with products that are unambiguously lacking in nutrition and easy to identify, like soda or candy. New York City, for example, attempted to ban sugar-sweetened beverages, and the state of Maine attempted to restrict sodas, candy, and any other taxable food items (Gundersen, 2015). The USDA overturned both restrictions.

Problems exist with “improving” the SNAP program by implementing even greater purchasing restrictions. First, there is no reason to believe that such a restriction would work. The restriction assumes that, under WIC-like requirements, households will substitute healthy foods for unhealthy foods when using SNAP benefits. What would most likely happen is that households would shift to purchasing unhealthy foods with cash. Second, such restrictions would likely lead to a drop in SNAP participation (Gundersen, 2015). Restricting choice is a paternalistic policy that would further stigmatize SNAP participation. It would give the impression that SNAP beneficiaries are assumed to have worse diets and that they cannot be trusted to make healthy food purchases. Participation would also drop due to increased transaction costs of purchasing items with SNAP. Not all stores would clearly mark which items were SNAP eligible nor should participants be expected to remember. The result would be longer, more frustrating shopping trips. Lastly, it is important to remember that for many SNAP recipients, freedom of choice is what makes the SNAP program popular and easy to use (Edin et al., 2013).

The most popular “improvement” was providing a monetary incentive to SNAP participants for purchasing healthy foods (Blumenthal et al., 2014; Leung et al., 2013). Mon-
etary (or financial) incentives, in this context, tend to be a rebate or voucher awarded to SNAP households for using their benefits to buy certain healthful foods, generally mineral-rich and nutrient-dense fruits and vegetables (i.e. leafy greens but not white potatoes). These monetary incentives for buying “targeted” fruits and vegetables (produce) are exclusive to SNAP participants. Much like a grocery stores loyalty card or a student ID card, retailers can “discriminate on price” (aka “target the incentive”) using SNAP Electronic Benefit Transfer (EBT) cards to identifying eligible participants. Monetary incentives in the food retail environment are popular for two main reason. First, the framing of the “improvement” is positive. Instead of “punishing” SNAP participants through paternal restriction or disincentives (not covered), monetary incentives reward participants for healthy shopping behavior (Gundersen, 2015). Retailers also prefer the positive framing of monetary incentives. For the moment, monetary incentives programs for SNAP participants are not wide spread. Taking up an incentive program, assuming the cost of implementation isn’t too expensive, creates an opportunity for retailers to differentiate themselves from their competitors (Hartmann, 2011). The second reason is a strong theoretical framework established by neoclassical economics supporting incentives as an effective mechanism for changing human behavior. In practice, however, incentives have had mixed results, but there is building evidence that incentives may work in the food retail space.

How, why, and to what effect incentives may encourage SNAP participants to buy more produce is the motivating question behind this paper.
1.1.1 Financial Incentives to Encourage Healthy Food Purchases

Economic theory would suggest that SNAP households are strong candidates for targeted financial incentives. SNAP households are highly aware of how much money they have (budget aware) and how much they are spending (budget conscious) (Wiig and Smith, 2009). SNAP itself functions under the economic theory that, by providing fungible cash for food, low-income households can shift money budgeted for food towards other expenses (Bartfeld et al., 2015). In practice, however, SNAP households treat EBT benefits differently, spending SNAP dollars differently from cash and other income sources—behavior better described by the concept of “mental accounting” (Hastings and Shapiro, 2018). This implies that, by further increasing the purchasing power of a household’s SNAP EBT benefits, the likelihood of those savings being shifted towards other expenses (e.g. car repairs, health care) is lower than economic theory would suggest. Therefore, under the assumption that SNAP households are fully informed about the DUFB incentive program, it would be reasonable to expect SNAP households to take advantage of a financial expensive that could double their monthly fruit and vegetable spending—and without substituting the saving towards other expenses.

The challenge, of course, is that many SNAP households do not have complete information. No households is fully aware of what deals or programs exists throughout any given store or supermarket chain at given time. And even if SNAP households were made fully aware of a specific incentive program, it does not imply the program will be easy to understand and to use. It is possible that, despite the potential savings, the program
is too burdensome or complicated to be worth the time and effort. Another challenge is the general unpopularity of fruits and vegetables. For some households, it can be that they are not aware of the benefits of eating more fruits and vegetables or a lack of knowledge of how to prepare them (Mancino and Guthrie, 2014). For others, it is more important to purchase items with a longer shelf life than produce (Leung et al., 2013). Therefore, while SNAP households appear in theory to be strong candidate for targeted financial incentives, we should temper our expectations given that learning and executing any program imposes a cognitive and temporal cost. And this would be especially true for less popular products like fruits and vegetables.

1.1.2 The Healthy Incentives Pilot

Research where SNAP participants are the target group is nascent. The USDA’s Food and Nutrition Services (FNS) ran the first large scale randomized control trial investigating the impact of a financial incentive for targeted fruits and vegetables in 2011. The experiment was called the Healthy Incentives Pilot (HIP). HIP is the major precursor to every incentive program currently being funding by the USDA. It also provides the data for the few papers recently published on incentives for SNAP participants.

A brief overview of the Healthy Incentives Pilot is necessary to provide context to, and contrast with, more recent financial incentive programs.

The USDA’s Food and Nutrition Services designed the Healthy Incentives Pilot. The pilot was funded by the Food, Conservation, and Energy Act of 2008 to test whether financial incentives would increase consumption of “targeted fruits and vegetables” (TFVs).
SNAP participants were the target group.

HIP was designed as a large scale randomized control trial (RCT). FNS partnered with the Massachusetts Department of Transitional Assistance to implement HIP. The pilot lasted from early 2011 to the end of 2012. The population included all 55,095 SNAP participants in Hampden County, MA. Hampden County is the poorest county in Massachusetts and has the highest rates of obesity and other diet-related chronic illness (e.g. type 2 diabetes).

Of the 55,095 SNAP participants, 7,500 were randomly assigned to the treatment group. The remainder fell into the control group. The treatment was a 30 cent (or 30%) rebate on every dollar spent on TFVs. The rebate was capped at $60 per month. To receive the rebate, selected SNAP participants had to use their EBT cards at participating retailers. The rebate, which was returned to their EBT account, could then be used on any food item. That is, the rebate could only be earned buying TFVs, but could be redeemed buying any SNAP eligible food item. Most HIP participants spent about $12 a month on TFVs, earning an average of $3.65 per month in rebates—drastically lower than the $60 per month rebate cap.

The evaluation was conducted using 24-hour dietary recall surveys. A total of 5,000 participants were selected to be surveyed, even split between treatment and control (2,500 HIP, 2,500 non-HIP). The first survey was conducted prior to the start of the pilot. This established a baseline. The second survey occurred 4 to 6 months in to the pilot and the third survey occurred 9 to 11 months in. (The variation, e.g. 4 to 6 months, was due to
The evaluation found that the 30% rebate lead to about a 26% increase in consumption of TFVs. This was equivalent to about 0.24 cups of TFVs. Roughly 60% of the increase was due to increased vegetable consumption and 40% due to increased fruit consumption. The effect, in absolute terms (0.24 cups), seems small. But a 0.87 price elasticity, relative to other results in the literature, is quite high—0.7 and 0.48, on average, for fruits and vegetables, respectively (Andreyeva et al., 2010).

Despite some limitations and technical problems—underreporting on the 24-hour recall survey and system glitches early in the pilot (see pages 60 and 208-210 of Bartlett et al. (2014))—HIP was considered to be an overall success (Klerman et al., 2014; Olsho et al., 2016). It implemented on of the largest, most complex RCTs to isolate how incentives can increase household consumption of TFVs. It also provided a feasible model for nationwide expansion (assuming cost reductions due to economies of scale; see An (2015)).

HIP also provides a framework for understanding how a financial incentive, expanded dramatically in one geographic area, could improve TFV consumption. But, as noted in the final HIP report, one of the most prominent retailers in Hampden County chose not to participate (page 61, Bartlett et al. (2014)). Its third-party processor decided it was too difficult and too costly to implement the financial incentive on its point-of-sale technology. This strategic behavior by the retailer, which had a significant presence in Hampden County, impacted where participants could use the incentive.

Most financial incentive programs work at the local level, expanding non-
randomly. We should anticipate certain retailers (firms) to behave strategically when participating in any of these incentive programs. Likewise, we should anticipate voluntary (non-random) self-selection by SNAP beneficiaries into these financial incentives programs. To this end, more research is needed to understand the impact of incentive programs under real-world conditions. HIP provided evidence that an incentive program can work, but barring state-wide or nation-wide adoption of point-of-sale financial incentives, we should expect the growth of any program to occur strategically and endogenously.

An example of such a financial incentive program for SNAP participants is the Double Up Food Bucks program (DUFB or Double Up). The non-random expansion and impact of this financial incentives program will remain the focus of this paper.

1.1.3 The Double Up Food Bucks Program

The success of HIP paved the way for the Food Insecurity Nutrition Initiative (FINI), established by section 4208(b) of the Agricultural Act of 2014 (aka 2014 Farm Bill). FINI—a 100-million-dollar initiative—in turn piloted numerous non-profit financial incentive programs aimed at improving the diets of SNAP participants.

Of specific interest is Double Up Food Bucks, an incentives-based program funded by FINI. In 2009, the non-profit organization Fair Food Network (FFN) launched the Double Up Food Bucks program in Detroit, Michigan. The intention of the program was to get more low-income families visiting and participating in local Detroit farmer’s markets. The mechanism for increasing participation was a dollar-for-dollar match of locally grown fruit and vegetable purchases. This subsidy was accessible only to low-income
families receiving SNAP benefits, who could exchange up to $20 of their benefits for a wooden token that could be used on up to $40 worth of locally grown produce.

The DUFB program was considered successful given it had expanded to more than 150 farmer’s markets in 2014 from just 5 farmer’s markets in 2009. SNAP benefits have been used more than 200,000 times to purchase fresh produce, with more than 10,000 first time SNAP customers visiting farmer’s markets in 2013 alone (Network, 2014). The program is considered by Fair Food Network to be a “three-fold” win given that the program helps local low-income families buy more fresh produce, provides new customers for local farmer’s, and stimulates the local food economy. Relative to farmer’s markets in other states, DUFB did seem to be bringing in substantially more SNAP dollars ($1.7 million in Michigan versus $307,000 in Illinois, the second largest).

A 5.17 million dollar FINI grant was awarded to Fair Food Network to help it pilot three adjustments to the Double Up Food Buck program (USDA NIFA, 2015). First, FFN needs to test DUFB as a year-round program in select locations instead of the current seasonal format. Second, shift away from the token system to providing DUFB electronically at point-of-sale. Third, the DUFB needs to expand from farmer’s markets into other retail environments, like supermarkets and grocery stores.

Successful expansion into supermarkets and grocery stores is critical. Approximately 80% of all SNAP benefits in 2015 were used in supermarkets or super stores (USDA FNS, 2016a). Less than 1% percent of SNAP benefits were used at local farmer’s markets. The amount of SNAP benefits used in local farmer’s markets has increased since 2009,
but no where near the growth necessary to reach the type of stores most frequented by low-income families. If localized financial incentive programs like DUBF are going to be considered one of the USDA’s many tools to increase food access and combat obesity, then they must be successfully implemented and scaled across supermarkets and grocery stores. Most importantly, incentive programs like DUBF must prove they are effective in changing purchasing habits within supermarket/grocery store food environments.

1.1.4 Double Up Food Bucks vs the Healthy Incentives Pilot

There are notable differences between DUBF and HIP that make the evaluation of DUBF more difficult. In short, HIP was implemented as an RCT. DUBF implementation is not. Let’s explore in greater detail.

HIP had substantially more participating stores, all within the same county (Hampden County, MA). DUBF has fewer participating stores, spread across many different counties, and across many different grocery store chains. Therefore, the probability of a SNAP shopper in Hampden County having walked into a HIP participating store was much higher than a SNAP shopper walking into any DUBF participating retailer.

The incentive delivery mechanisms also differ. First, all SNAP beneficiaries who shop at a DUBF participating store receive the benefit automatically. In other words, SNAP households that patron a store assigned to the DUBF program receive the incentive regardless of their intentions or awareness of the DUBF incentive. In contrast, SNAP households assigned to the HIP treatment group were made aware of incentive and only treated households were eligible to use it (even if they didn’t quite understand how the incentive program
 worked — see Bartlett et al. (2014)). Households in the control group were not aware of the incentive and were not eligible to use it. In other words, stores were assigned to the “treatment” group in the DUFB intervention while households were assigned to the “treatment” group in the HIP intervention.

Second, the DUFB financial incentive is substantially larger but more restrictive. The DUFB incentive is a dollar-for-dollar match of locally grown produce purchases capped at $20 per day. The matched dollars are accrued as points on a store loyalty card. Existing points are then automatically redeemed as dollars on any fresh produce purchases, not just locally grown produce. In comparison, the HIP financial incentive was a return of 30 cents per dollar spent on TFVs which could be spent on any food item. That is, the DUFB incentive doubles the purchasing power of every dollars spent on TFVs only for more TFVs; the HIP incentive increased the purchasing power of every dollars spent on TFVs by 30% for any SNAP eligible food item.

Finally, the experimental design of HIP allowed researchers to form a causal interpretation of their results; the average treatment effect is the same as the average treatment effect on the treated. Any difference in the purchase and consumption of TFV between the treatment and control groups could therefore be attributed directly to the incentive. Measuring causality for the DUFB program is less straightforward, requiring more sophisticated econometric methods. However, analyses of RCTs like HIP are the exception, not the rule. How DUFB and similar financial incentive programs are implemented and analyzed is the norm. The contribution of this paper will be evaluating and understanding the impact of
DUFB, given that DUFB and similar programs are implemented in the “real-world” under, at best, quasi-experimental conditions.

1.1.5 Evaluating Double Up Food Bucks

DUFB’s expansion and implementation into supermarkets and grocery stores did not follow standard experimental design. Fair Food Network searched for local partners willing to participate in DUFB. Not all grocery stores, especially the smaller independent stores, had the capacity to implement the point-of-sale technology necessary for the incentive—even if FFN offered to help cover the upgrade costs. The result is a self-selected grocery retailer participating in DUFB. This, in some ways, parallels what occurred in HIP, where one of the largest retailers decided integrating their point-of-sale systems to include the incentive was too expensive. This type of strategic firm behavior is important to consider, even if complicates the evaluation of an incentive program like DUFB.

In reality, grocery retailers seek to maximize profits and will opt to participate only if they expect to profit. Similarly, individuals will self-select into participating; participation is optional and more likely to occur with well-informed and motivated SNAP shoppers. Selection, in this case, is a feature, not a flaw, of such incentive programs when implemented by non-profits or policy makers. The evidence, thanks to HIP, exists that incentives can lead to an increase in consumption. The goal of this paper is therefore to accurately measure the effect of the DUFB on produce spending while controlling for self-selection.

Fair Food Network started testing and gathering data from grocery stores imple-
menting DUFB in 2014. One of FFN’s largest partners, a large grocery retail and distribution company, piloted the program in 2 of its stores in 2014. The company expanded to 5 stores in 2015 and then to 17 of 62 stores in 2016. Rapid scaling was possible due to the point-of-sale technology used by the company to implement DUFB across its stores. It is a useful case study of what happens when a firm strategically scales DUFB across numerous grocery stores that span different geographic areas and populations.

Transaction data was provided for every store that has, at any point, participated in DUFB program from 2014 - 2016. These data are complete (i.e. no records have been removed) and at the item level. A complete set of data was also be provided from another 15 stores where DUFB was not implemented. Only the 2016 data was used in the evaluation of the incentive. The reasons are explained in the proceeding chapters.

Currently, limited research exists evaluating targeted financial incentives using a complete set of store transaction data. HIP, for example, only had transactions records for SNAP EBT cards. Transactions, should a different tender be used by the same individual, could not be observed. Therefore, these data provide an unprecedented opportunity to analyze how the DUFB financial incentive performs under real-world conditions. This paper performed one of the first evaluations of a financial incentive, targeted at SNAP participants, using a complete set of data, from multiple stores, collected under real-world conditions.

1.1.6 Hypothesis

My original hypothesis, considering the literature about financial incentives and the spending and food preferences of most US households (SNAP or otherwise), was that the DUFB
financial incentive would have a small but measurable impact on produce spending. I ex-
pected to find produce spending having increased by about 1 to 2 percent for SNAP house-
holds that frequented stores assigned to the DUFB program. I expected the difference to
fade during the duration of the program, disappearing once the incentive was no longer
available. The results, as I will show below, are in line with my original hypothesis.\textsuperscript{3}

1.1.7 DUFB Mechanisms

DUFB incentive mechanisms can differ. Existing mechanisms fall into three categories:
earn/redeem DUFB points via loyalty card, single-use paper coupon, and immediate dis-
count. The earn/redeem DUFB point system was used by the retailer that provided these
data. For information on the other two DUFB mechanisms, see (Margaret Schnuck, 2016).

The DUFB mechanism for the grocery store chain that provided the data for this
paper was a point system. SNAP households could earn points by buying locally grown
produce using their SNAP EBT card and their loyalty card. Earning DUFB points required
using a loyalty card. The retailer used the cards to keep track of points. Each dollar spent
buying locally grown produce earned a DUFB point. SNAP shoppers were eligible to re-
ceive up to 20 dollars worth of DUFB points (20 points) per day. Earned points were not
immediately reflected on loyalty cards but were redeemable the day after.

Shoppers could redeem points on any eligible produce (excludes frozen and
canned fruits and vegetables). A loyalty card, given it tracked points, was required to

\textsuperscript{3}A record of my original hypothesis can be found in my prospectus, first logged in Feb
2017. This was 6 months before receiving the data (August 2017). The commit record
of my prospectus can be found at \url{https://github.com/dantonnoriega/prospectus-draft/blob/}
\url{6e48aa56dcbeb36e5c366516759c3feca3b5adbb/01.20-question.Rmd}.
redeem points. redeeming points, however, did not require spending SNAP EBT benefits; redeeming points was possible using any method of payment (tender).

This earn/redeem DUFB point system of the participating grocery retailer had a few important details. Already mentioned, but worth reiterating, is that it took a day (24 hours) to process earned points. This forced SNAP shoppers to delay the reward of the DUFB incentive earned by at least a day. While perhaps a technological necessity (it takes time to process and reflect earned points on loyalty card accounts), this delayed transactional utility for the SNAP shopper (Thaler, 1985). Delaying transactional utility can be expected to reduce the “pleasure” and effectiveness window of the DUFB incentive.

Second, the incentive alternated between two states: earning and redeeming. A loyalty card with a DUFB point balance of zero is in an “earning” state. After earning DUFB points by buying locally grown fresh produce, the card switches to a “redeeming” state; loyalty cards with a point balance greater than zero will redeem until the point balance is once again zero. This is to discourage shoppers from strategically “banking” earned points. Some shoppers, for example, may want to “bank” points then strategically redeemed them all at once (say for a holiday shopping trip). The closest a SNAP shopper could come to “banking” points is by spending a lot during a single transaction, getting as close to the 20 point cap as possible. This could happen, in theory. SNAP shoppers tend to spend benefits in one large shopping trip soon after receiving monthly benefits (Wilde and Ranney, 2000). These shopping trips correspond to the single largest opportunity to earn DUFB points.

Third, the points earned were not directly communicated to the shopper before or...
during the moment of sale. In other words, the fact that shoppers were earning points was not salient. No clear feedback connected buying fresh healthy produce to earning points. Earned points (those processed in prior shopping trips) were only communicated via receipt (printed at the bottom) 24-hours after the fact. The other option was to check on-line or visiting an in-store kiosks. But no earning information was shared with the shopper during the current sale. Earning points, therefore, end up being feeling like any other shopping trip. Certainly a bell shouldn’t ring when SNAP shoppers earn points. One of the most positive consequences of moving to an EBT card is that the potential stigma of using food stamps has greatly diminished. But shoppers could benefit from some sort of salient and immediate feedback that is informative without producing a spotlighting or stigmatizing effect. For example, shoppers could be told, “You saved $4.50 today and you also earned 5 DUFB points”. Redeeming points, however, was better communicated. After any transaction where DUFB points were redeemed, each point spent was displayed as line item discount on the customer’s receipt (see Figure 1.1). This provided one of the best opportunities for uninformed SNAP households to learn about the program.

Lastly, it was not obvious to SNAP shopper which produce items were eligible for earning points. Earning points was only possible on locally grown produce. According to a representative of the grocery retailer, the tagging of local produce was generally done at the discretion of store management. Only a few products were tagged by corporate. It was possible for an item to maintain the same UPC code yet change from a local supplier to an international one, the former being eligible for earning points. This is confusing.
Paired with the fact that knowing whether points were earned can take more than 24 hours, this could limit opportunities for SNAP shoppers to learn how to maximize the incentive. Redeeming was easier to determine given it essentially covered all produce items.

**Salience, Complexity, and Fading Effect**

Financial incentives have a long history of mixed results. Financial incentives have been shown to help individuals commit to regular exercise, improve dieting, increase weight loss, and to quit smoking, but the intended effect of the financial incentives were often short-term, disappearing once the incentive was no longer available (see Gneezy et al. (2011) and Cawley (2015) for an overview). But there is nothing irrational about no longer responding to an incentive that disappears aka “temporary compliance” (Kohn, 1998). What traditional economic theory fails to explain, however, are actors responding with fading intensity to the incentive even while in available.

To explain why a fading effect of the incentive for the duration of the DUFB program was hypothesized, it is necessary to turn to concepts from behavioral economics. Simple and salient interventions can often out perform more generous, yet more complex, interventions (Diamond and Vartiainen, 2012). Complexity introduces cognitive costs that affect participation. The mechanism used to implement the DUFB financial incentive was simple to use. SNAP shoppers used their loyalty card and EBT benefits like any other transaction. But there were aspects of the DUFB incentive program independent of the mechanism that were not simple nor salient. As mentioned, identifying which produce items earned points was not straightforward. Likewise, determining how many points were earned
Figure 1.1: Example of Customer Receipt with DUFB Points Redeemed
in a transaction was not salient. Learning how the DUFB program worked was therefore a type of cognitive cost. So was learning how to make the most of the DUFB program from shopping trip to shopping trip. It was easy imagine these costs being constantly weighed towards the benefits of the DUFB incentive. For some SNAP shoppers, especially those less interested in produce, these costs quickly surpass any benefits of the DUFB program. They would, after trying out the DUFB incentive program, lose interest and regress back to their normal shopping routine. This, it was hypothesized, would be observed as a month-to-month fading effect on produce spending, positive at the start of the DUFB program and fading to zero by the end of the program.

**Identification and Self-Selection into the DUFB Program**

How a participant responds to assignment is generally referred to as “compliance” (Angrist and Pischke, 2008). But in this case, the stores, not the individual shoppers, have been “assigned” to a treatment or control group. Stores, if assigned to the treatment group by the retail chain, are “compliers” by default; the store’s point-of-sale (POS) system is altered to implement DUFB. Some stores (4) behaved somewhat like “always-takers”, having asked to participate in DUFB, but most store (13) are “compliers”.

How, then, does one think about SNAP households participation in the DUFB program if it is stores that are ultimately assigned to the DUFB program? SNAP shoppers have the option to benefit from the program without ever being “assigned” to any treatment

---

4 Participants can be further categorized into “compliers”, “never-takers”, “always-takers”, and “defiers”. These categorizations provide useful terminology but are not relevant in the context of the DUFB incentive.

5 While this creates some worries of “self-selection” by stores, this bias is handled by a model that includes a store-level fixed-effect.
group. A shopper’s participation in DUFB is therefore also driven by self-selection. Interest in the DUFB incentive may depend on unobservable variables corresponding to individual shoppers, stores, and the retail chain. For example, demographics, price sensitivity, food preferences, health consciousness are all unobservable variables that could affect shopper DUFB activity. Other unobservable variables include how effectively the retail chain markets the DUFB program to the management of participating stores and how effectively this information is then relayed to individual shoppers. Management’s enthusiasm for the program is likewise a unobservable retail chain and store-level variable.

But identifying whether some SNAP households were responding to the DUFB incentive depended on enough households self-selecting into participating in the program. If it was assumed that shoppers in DUFB stores were relatively similar to shoppers in non-DUFB stores, then observing an increase in produce spending required the existence of SNAP households actively choosing to participate in the DUFB program. In other words, identification of the DUFB program’s impact on produce spending depended on SNAP households interested in buying more produce self-selecting into the program, differentiating themselves from SNAP shoppers in the non-DUFB stores. Crucially, SNAP households that were unresponsive to the incentive in DUFB stores were assumed “canceled out” by SNAP households in non-DUFB stores. In short, any measurable increase in produce spending required there be enough SNAP households self-selecting into using the DUFB program, buying more produce than they would have otherwise purchased.
DUFB Incentive Inconsistency Across Years

How the DUFB incentive worked in 2016 is distinct from 2014 and 2015. The DUFB incentive in 2016 worked by earning points for each dollar spent on locally grown fresh produce.\(^6\) Points were then redeemed automatically on any fresh produce. However, in 2014 and 2015, the incentive was the opposite. In those two years, the DUFB incentive worked by earning points on any fresh produce, automatically redeeming points on locally grown fresh produce.

This is important because locally grown fresh produce is a much smaller subset than any fresh produce. Therefore, in years 2014 and 2015, shoppers could easily earn points but had a constrained set of produce on which to redeem points. The opposite was true in 2016. Because of this, the 2014 and 2015 data should not be pooled with the 2016 data to estimates the effect of the DUFB incentive. Modeling must be done separately for each year given changes in the implementation mechanism.

1.2 Data

The data for this analysis were provided by a large US grocery distributor and retailer. The retailer has partnered with Fair Food Network since 2014, when it first piloted the DUFB program in 2 stores. DUFB was implemented in only one of many grocery chain operated by the retailer. The grocery chain that provided the data has more than 60 stores.

The retailer expanded the DUFB program to 5 stores in 2015 and then to 17 stores

\(^6\)Recall that each point is equal to one dollar.
in 2016. Data are available for these three years—2014 through 2016—from 46 stores. This paper, however, will perform an analysis using only the 2016 data. The reasons are detailed in a later section, but in brief, there was greater balance between the number of stores implementing DUFB in 2016 than not, and how the DUFB incentive program worked also differed in 2016 versus 2014 and 2015. Of the 46 stores with available data, 29 stores did not implement DUFB in 2016. These 29 stores serve as “controls” for the 17 stores that did implement DUFB in 2016. The quotes here signify that these are reference terms. The terminology is somewhat misleading; the use of “treatment” and “control” could lead one to think store assignment was random. It was not. Store assignment is also detailed in a section below.

The data are transaction-item level. Every row of data denotes an item within a single transaction. “Transaction” refers to any visit to the cash register where a customer pays or returns an item. For each row of transaction-item data there is a loyalty card number, a store identifier, register ID, transaction ID, date and time of purchase, payment type (aka tender), item universal produce code (UPC), an item description, item department, discounts used, listed price, paid price, quantity or weight, and how many, if any, DUFB points were earned or redeemed. No demographic data were made available for privacy reasons. All transactions can be uniquely identified via transaction (“trx”) date, transaction time, store ID, transaction ID, and register (terminal) ID.
1.2.1 Loyalty Card Concerns

What makes these data valuable are the loyalty card numbers. Loyalty cards help link transactions to the same identifier over time, producing an (unbalanced) panel. Without such an identifier, estimating the impact of the DUFB incentive program would be extremely difficult. Having data with loyalty card numbers, however, does not guarantee significant results when measuring the DUFB treatment effect. It is necessary that loyalty card usage be ubiquitous across users, specifically within the populations of interest—SNAP participants.

There are two general concerns with these type data as it pertains to loyalty card numbers. As mentioned, the first is the possibility that a significant proportion of transactions are not linked to loyalty cards. This would be especially concerning if loyalty card usage was systematically unpopular within the subpopulation of interest, SNAP participating households. If loyalty card usage does not account for the vast majority of purchases by likely SNAP participants, then it will be difficult to claim similarity between SNAP populations across DUFB and non-DUFB stores. Fortunately, in these data, loyalty card usage is very popular with likely SNAP participants (above 90%). Dropping non-loyalty card transactions should not affect results.

The second concern is that it is not possible to determine with certainty that a single household is consistently using the same loyalty card. Grocery store chains share the same concern given it makes learning the behavioral patterns of households more difficult. Efforts were made to remove obviously problematic loyalty card numbers from these data. “Obviously problematic” loyalty card numbers were (1) observed being used multiple times
per day and almost every day of the year or (2) were associated with thousands of dollars above median monthly sales of loyalty card customers by store. These were assumed to be loyalty cards scanned by store employees to offer discounts to customers who did not have, or forgot, their loyalty cards at purchase. No other transactions were removed from these data. The driving assumption was that, should multiple households use the same loyalty card number, and after the removal of problematic card numbers, these “violations” occurred with similar frequency across all stores and would be difference out. There is no way to prove this assumption given the data.

**Loyalty Card Usage**

**Table 1.1:** Pooling Together all Transactions and Items Purchased. Grouped by Year.

<table>
<thead>
<tr>
<th>Year</th>
<th>% Loyalty Trxs</th>
<th>% Loyalty Items</th>
<th>Total Trxs</th>
<th>Total Items</th>
<th>Num. Stores</th>
<th>Num. Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>81.04</td>
<td>95.06</td>
<td>15633723</td>
<td>131681498</td>
<td>46</td>
<td>8</td>
</tr>
<tr>
<td>2015</td>
<td>81.01</td>
<td>95.05</td>
<td>22433377</td>
<td>186778717</td>
<td>46</td>
<td>12</td>
</tr>
<tr>
<td>2016</td>
<td>81.16</td>
<td>95.07</td>
<td>22502710</td>
<td>185164744</td>
<td>46</td>
<td>12</td>
</tr>
</tbody>
</table>

**Table 1.2:** Pooling Together all SNAP Transactions and Items Purchased. Grouped by Year.

<table>
<thead>
<tr>
<th>Year</th>
<th>% Loyalty Trxs</th>
<th>% Loyalty Items</th>
<th>Total Trxs</th>
<th>Total Items</th>
<th>Num. Stores</th>
<th>Num. Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>91.73</td>
<td>97.20</td>
<td>1040198</td>
<td>13078472</td>
<td>46</td>
<td>8</td>
</tr>
<tr>
<td>2015</td>
<td>91.76</td>
<td>97.23</td>
<td>1335647</td>
<td>16805052</td>
<td>46</td>
<td>12</td>
</tr>
<tr>
<td>2016</td>
<td>92.19</td>
<td>97.38</td>
<td>1208266</td>
<td>15100941</td>
<td>46</td>
<td>12</td>
</tr>
</tbody>
</table>

These data contain hundreds of millions of items purchased across tens of millions of unique transactions per year. Consistently, year over year and pooling across all stores, loyalty
cards were used in 81% of all transactions and 95% of all items purchased. Loyalty card usage rates are even higher when SNAP dollars were used to purchase part or all of a transaction (aka “SNAP Trxs”), increasing to roughly 92% and 97% respectively. See Tables 1.1 and 1.2 for more specific values. Note that sales data were intentionally excluded at the request of the company.

This is a very high level summary. Pooling all transactions together by year fails to capture the substantial variability that exists at the store level.

Below are tables of monthly aggregates for each store, pooled by year and, when specified, by transaction type (SNAP versus Non-SNAP). There are 32 complete months of data from 46 stores totaling 1472 store-month values (8 months for 2014, 12 for 2015 and 2016). Table 1.3 displays the average number of loyalty card transactions observed per month, per store, and by type (SNAP Trx means SNAP dollars were used to purchase part or all of a transaction). The table also displays the average of loyalty card transactions per month and per store as a percentage of all observed transactions. As expected, SNAP dollars are not used for the vast majority of loyalty card transactions (roughly 5 to 6 percent of all transactions).

Table 1.3: Loyalty Transactions, SNAP vs Non-SNAP (Store-Month by Year). Percentages calculated out of all observations.

<table>
<thead>
<tr>
<th>Year</th>
<th>SNAP Trx</th>
<th>Avg. Loyalty Trxs</th>
<th>SD. Loyalty Trxs</th>
<th>Avg. Loyalty %</th>
<th>SD. Loyalty %</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>No</td>
<td>31834.18</td>
<td>10075.61</td>
<td>76.08</td>
<td>6.91</td>
</tr>
<tr>
<td>2015</td>
<td>No</td>
<td>30701.37</td>
<td>10438.12</td>
<td>76.76</td>
<td>6.68</td>
</tr>
<tr>
<td>2016</td>
<td>No</td>
<td>31069.05</td>
<td>10990.25</td>
<td>77.47</td>
<td>6.30</td>
</tr>
<tr>
<td>2014</td>
<td>Yes</td>
<td>2592.99</td>
<td>1291.22</td>
<td>6.38</td>
<td>2.65</td>
</tr>
<tr>
<td>2015</td>
<td>Yes</td>
<td>2220.37</td>
<td>1142.38</td>
<td>5.80</td>
<td>2.52</td>
</tr>
</tbody>
</table>
Table 1.4: Loyalty Transactions, Pooled

<table>
<thead>
<tr>
<th>Year</th>
<th>Avg. Loyalty Trx</th>
<th>SD. Loyalty Trx</th>
<th>Avg. Loyalty Trx %</th>
<th>SD. Loyalty Trx %</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>34427.17</td>
<td>10524.12</td>
<td>82.46</td>
<td>6.94</td>
</tr>
<tr>
<td>2015</td>
<td>32921.74</td>
<td>10813.99</td>
<td>82.55</td>
<td>6.79</td>
</tr>
<tr>
<td>2016</td>
<td>33087.08</td>
<td>11356.47</td>
<td>82.74</td>
<td>6.52</td>
</tr>
</tbody>
</table>

Table 1.4 pools all transactions together—that is, without conditioning on whether SNAP tender was used to purchase all or part of a transaction. Pooled together, loyalty cards are used in around 82.58 percent of all transactions. This a large fraction all transactions observed. The above claim that dropping transactions without loyalty card numbers “should not affect the results” is in reference to this high loyalty card usage observed across all transactions. Note that this is essentially the same information as Table 1.3 where one would get the same percentages after summing the column Avg. Loyalty, % of All Trx grouped by year. However, Table 1.3 differs from Table 1.1 given that the values in Table 1.3 are computed by aggregating at the store-year-month level before summarizing by year.

Table 1.5: Loyalty Transactions, SNAP Only

<table>
<thead>
<tr>
<th>Year</th>
<th>SNAP Trx</th>
<th>Avg. Loyalty Trx</th>
<th>SD. Loyalty Trx</th>
<th>Avg. Loyalty Trx %</th>
<th>SD. Loyalty Trx %</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>Yes</td>
<td>2592.99</td>
<td>1291.22</td>
<td>92.27</td>
<td>3.84</td>
</tr>
<tr>
<td>2015</td>
<td>Yes</td>
<td>2220.37</td>
<td>1142.38</td>
<td>92.37</td>
<td>3.54</td>
</tr>
<tr>
<td>2016</td>
<td>Yes</td>
<td>2018.03</td>
<td>1074.53</td>
<td>92.71</td>
<td>3.12</td>
</tr>
</tbody>
</table>

Table 1.5 are results conditional on the subset of non-SNAP transactions only. We
get similar results as Table 1.4. This is to be expected; most transactions do not involved SNAP tender and the average should converge closely to the pooled average given the small sample size of SNAP transactions (relative to all transactions).

Figure 1.2: Loyalty Card Transactions as Percentage of all Transactions (Histogram)

Figure 1.2 is a histogram where each observation (count) is the percentage of loyalty card transactions observed in a single month from a single store. There are 32 observations (months) per store. Each store is assigned a unique color (some colors are similar enough that it can be hard to distinguish). The color clusters help show how each store tends to have a relatively stable percentage of loyalty card transactions from month to month. Ignoring color, the overall histogram shows the global variation and expected percentage value should one choose a store at random and attempt to infer the loyalty card percentage.
Figure 1.3: Loyalty Card Transactions as Percentage of ALL Transactions (Density)

Figure 1.3 helps visualize how usage loyalty rate vary across stores but also shows the relative stability of usage rates within each store. This takes the chaotic colors of the histogram and groups the colors into separate clusters. The more overlap in the density plot, the higher the bars would be in the histogram plot.

Figure 1.4 help visualize the stability from year to year conditional on DUFB assignment. It also displays how within-store loyalty card usage rates remain stable across across time; there are no dramatic shifts left or right in the densities of the top nor bottom panels. The density plots do not perfectly overlap, but noise is expected when plotting of a few points per year.

Figure 1.5 shows how, once one subsets to SNAP participants only, the distributions shift up about 10 percentage points regardless of DUFB assignment. The important
Figure 1.4: Loyalty Card Transactions as a Percentage of Total Monthly Transactions

Figure 1.5: Loyalty Card Transactions as Percentage of SNAP Transactions
takeaway is that, for the population of interest (SNAP participants), loyalty card usage is always higher than the broader population. In other words, if a customer is also a SNAP participant, they are much more likely to use their loyalty card while shopping no matter the store being visited.

One may be tempted to infer something about the difference in pooled means between groups by DUFB assignment. One should not. First, the difference in means is likely the result of sampling error; there are few stores within each group. Second, and more importantly, the pooled difference in means doesn’t tell us anything about the effect of the DUFB. More sophisticated methods are required to measure the effect of the DUFB program given that DUFB assignment was not random. The point of the plots is to show how, conditional on SNAP participation, the probability of a customer using a loyalty card increases by 10 points. Furthermore, the plots highlight that within a specific store, loyalty card usage is relatively stable over-time, clustering around the same range of values.

1.2.2 Gradual Decrease in SNAP Transactions

SNAP loyalty card transaction as a percentage of all transactions decreases gradually over time. This would be concern only if it was driven by something that affected some stores but not others. There is no evidence that is the case; all stores see a slight decrease in SNAP purchases from 2014 to 2016.

More likely the data are reflecting a decrease in overall SNAP enrollment over the same time period (see plot 1.6). SNAP enrollment negatively correlated with improving economic conditions. And broader economic trends impact all stores independent of DUFB
assignment.

Figure 1.6: Month to Month US Household SNAP Enrollment, 2012 - 2017.

1.2.3 Overview of Store Selection and Expansion

How the 17 “treatment” stores and the original 15 “control” stores were selected in 2016 is important. First and foremost, selection was not random. Stores were either selected by the company (13 of 17) or self-selected into DUFB (4 of 17). Second, the original 15 control stores were selected after the selection of the 17 treatment stores. Data from all remaining stores was requested but the request was denied; only 15 stores had been approved by the company’s management. Understanding the origins of the 17 treatment stores had implications for how, and ultimately which of, the original 15 control stores were selected.
1.2.4 Selection and Expansion of DUFB Stores

The first 2 stores were piloted with DUFB in 2014. Both were in geographically distinct areas (these will be referred to as “Node 0” and “Node 1”). There was a small expansion adding 3 more stores in 2015. The 3 stores were selected because they were geographically close to the 2 original pilot stores (2 close to Node 0, 1 close to Node 1). The 5 stores are referred to as the “core”. The location of these 5 stores, separated in two clusters, established the geographic constraints that were then used to determine most of the additional stores in 2016.

DUFB was expanded to 12 more stores in 2016, totaling 17. Of those 12, 6 were selected due to their proximity to the 5 core stores, their SNAP EBT\textsuperscript{7} sales figures, and similarity in surrounding demographics (high population density, more African-American). In other words, 9 of the 17 stores—excluding the initial 2 pilot stores—were selected on a set of observable characteristics. The remaining 6 stores were not.

Of the remaining 6 stores, 4 asked if they could be included in the program. These stores self-selected into DUFB, making these stores fundamentally distinct. They were considered, and then included, only because they fell within the “Top 50”. The final 2 stores were selected by the company for “strategic business decision”. The best interpretation of this is that the company thought that DUFB would provide a competitive edge to the 2 included stores given some internal calculus. How the company came to this decision is unknown and therefore unobserved.

\textsuperscript{7}Electronic Benefit Transfer.
Table 1.6: Year by Year DUFB Store Selection

<table>
<thead>
<tr>
<th>Store</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>pilot</td>
<td>pilot</td>
<td>pilot</td>
</tr>
<tr>
<td>2</td>
<td>pilot</td>
<td>pilot</td>
<td>pilot</td>
</tr>
<tr>
<td>3</td>
<td>assigned</td>
<td>assigned</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>assigned</td>
<td>assigned</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>assigned</td>
<td>assigned</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>assigned</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>assigned</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>assigned</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>assigned</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>assigned</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>assigned</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>self-selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>self-selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>self-selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>self-selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>unobserved</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>unobserved</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Stores 1 and 2 represent the initial 2014 pilot stores.

Table 1.6 helps understand the year by year expansion of DUFB. Stores are classified as either assigned, self-selected, or unobserved. To be assigned means a store’s participation in DUFB was determined (assigned) by the company; self-selected means the store asked the company to participate; unobserved means that the company selected the store to participate in DUFB but for unknown and unobserved reasons. Numbers were assigned to each store for easy reference but otherwise have no meaningful interpretation.

1.2.5 Expansion on Observables

An example expansion on observables (using fake data) can be seen in Figure 1.7. Blue dots denote stores with DUFB, pink dots denote without. In the top frame, one can see two blue dots. These blue dots simulate the first two pilot stores in 2014. The left blue dot is Node 0 and the right blue dot is Node 1. The gray zones represent areas of higher population density.
Dark gray is considered urban, defined as having a population density of 1500 persons or more per square mile. The light gray are small towns and cities, more densely populated than very rural areas, but could not be considered urban. The expansion in 2015 (middle frame) proceeds to the stores closest to the original pilot stores. The expansion continues to 6 more stores in 2016 (bottom frame) away from the nodes but also along areas of higher population density.

Not conveyed in Figure 1.7 is that the 2015 and 2016 expansions also move through stores that happen to be “highly ranked”—that is, have relatively higher SNAP EBT sales.\textsuperscript{8} Also not conveyed is the fact that there is a strong correlation between geography, population density, racial composition, and SNAP EBT sales. The 2015 expansion to the most nearby stores also meant that it was an expansion to stores with high SNAP EBT sales in densely populated, African-American neighborhoods. The 2016 DUFB expansion was more explicit given that set of feasible stores substantially increases as one moves away from each node. DUFB stores were thus specifically selected not just by geographic proximity, but also by SNAP EBT sales ranking and demographic compositions similar to the initial 2014 stores.

**Expansion Data**

Data for about each store was built by merging 4 different sources. The core data came from the grocery retailer directly, which provided a list of stores participating in DUFB from 2014 - 2016. The grocery retailer also provided a list of stores ranked by EBT sales as

\footnote{\textsuperscript{8}All stores within the chain were ranked by SNAP EBT sales as a percentage of total sales.}
a fraction of total store sales and the size (square footage) of each store. Demographic and socioeconomic data came from the Data Science Toolkit API\(^9\) (DSTK) and the American Communities Survey API\(^10\) (ACS). The DSTK API provides access to US Census data from 2000 at the census block level and the ACS API provides data spanning 2010 - 2014 at the zip code level. Lastly, data was extract by mining the website of the participating grocery chain.

Matching was done with the ACS data. The ACS zip code data was preferred because it provided income and housing data. Zip code level demographics are sufficiently descriptive; stores are evenly distributed across zip codes. Specifically, 58 stores are spread across 58 zip codes and 4 stores split between 2 zip codes (60 zip codes and 62 stores).

Ideally, prior to matching, demographic data from the neighborhoods surrounding the store, who shopped at the store, and how the store was performing, its size, and goods made available would be known. Unfortunately, most of the publicly available data was not store-specific. The only store-specific data came either from the retail parent company directly or from scraping the company website.

### 1.2.6 Selection of Control Stores

Ideally, all remaining stores would have been available to use as a control group but initially the parent company only approved data be released for 15 stores. This left the added—and incredibly important—step of selecting the control stores since the company approved, but did not explicitly select, the 15 stores.

\(^9\)http://www.datasciencetoolkit.org/
Under this initial 15 store constraint, the next best option would have been for the parent company to provide a sample of transactions—perhaps a week or so—for every store, giving the option to dig through these data and determine the 15 best matches to the 17 DUFB (“treated”) stores using detailed and observable data. The parent company did not want to do this.
As noted above, the parent company provided a list of the grocery chain’s stores that included store address and internal estimates for SNAP dollar spending as a percentage of all spending. Upon request for more data, estimated number of store employees and the physical size (in square footage) of each store were provided by the parent company. These data, plus anything I could find publicly available about each store, were what was initially available to inform the selection of the 15 control stores.

Selecting the control stores proceeded in two steps. First, stores that either self-selected or were selected using some unobservable criteria were matched using Coarsened Exact Matching (CEM) (Iacus et al., 2011a). Second, stores assigned DUFB were pooled with nearby control stores and then scored using a linear probability model. Each step is summarized below.

**Step 1: Coarsened Exact Matching**

The 6 stores classified as self-selected or unobserved (stores 12 through 17; see Table 1.6) were compared against all possible control stores for matches. Matching was done across 5 dimensions: race, income, population density, store attributes, store EBT sales. One variable per dimension was selected: percentage of population that is African-American (zip code level); people per square mile (zip code level); median income for people who have received SNAP or similar assistance (zip code level); the number of associates employed in each store; and the percentage of total stores sales attributed to EBT/SNAP.

Of the 6 stores (stores 12 - 17), only 3 produced viable matches. However, each of the 3 matched stores had matched to more than one control stores. The closest stores,
by driving distance, were selected as the tie-breaker for each matched store. Stores were sufficiently far apart, with very sparsely populated areas between, that “spill-over” was considered unlikely. That is, it is considered unlikely that a shopper near a store without DUFB would opt to drive 30 or more minutes to shop at the store with DUFB.

This left 12 stores to be allotted to the control group and 3 treatment stores to be effectively discarded.

Step 2: Scoring via Linear Probability Model

Assignment to treatment and control can be close-to-perfectly determined since we know and observe the criteria used for assignment: geographic distance from an initial store (node), SNAP EBT sales rank, and demographics—specifically population density and percentage African-American.\(^{11}\) A scoring function was created by fitting a linear probability model to all stores within 140 kilometers of the two initial pilot stores.

\[
s = P(D = 1 | X, N)
\]

\[
= X\hat{\beta} + \alpha N + (X \odot N) \gamma
\]

\(s\) are the fitted values of the estimated linear probability model; \(D \in \{0, 1\}\) is a \(n \times 1\) vector of store assignments to DUFB; \(X\) is an \(n \times k\) matrix of normalized observable covariates that determine assignment; \(N \in \{0, 1\}\) is an \(n \times 1\) dummy vector denoting the closest pilot store aka “Node”, where 0 is Node 0 and 1 is Node 1. \(\odot\) represents element-

\(^{11}\)It should be noted that the company did not explicitly say population and race were part of the selection criteria. Instead, they said something along the lines of “stores serving a similar population as the original stores.”
wise multiplication aka “Hadamard product”.

Stores were sorted by the fitted values of the model, $s$. There is perfect separation between DUFB stores and those without (see Figure 1.8). Therefore, the top 11 stores by score value are all DUFB stores. The next 12 stores by score value are then allotted to the control group.

![Figure 1.8: Store Score vs DUFB Assignment](image)

### 1.2.7 An Additional 14 Stores

In 2017, after the initial 15 control stores had been selected, data for an additional 14 stores was provided retroactively for 2014 - 2016. The DUFB program had been expanded in 2017 to include the 15 stores initially selected as part of the control group, plus an additional 14 stores the parent company had selected for reasons not explained.

These additional stores bolstered the 2016 sample size of the control group to
almost double that of the treated group (29 control vs 17 treatment). Furthermore, the 14 additional stores all came from the same rural geographic area as the 6 stores that had either self-selected or were included from some unknown reason in 2016. These 14 stores essentially helped bolster the most underrepresented set of stores in the control group—the 3 rural stores—and provided a much needed counterbalance to the 6 rural stores in the treatment group.

Despite this stroke of luck and generosity, it is important to highlight that it was not possible to determine the quality of the selection procedure until after receiving the transaction data. Prior to receiving transaction data for any of the stores—whether it was initial 32 or the eventual 46—I did not know if the treatment and control groups would be balanced across features. While not necessarily detrimental to the eventual method of analysis (difference-in-difference-in-differences), it helps build a general sense of comfort that there is nothing drastically different between the two groups.

Other important assumptions required to measure the impact of the DUFB incentive, such as the “common trends” assumption, can be better defended if there is relative parity between the two groups prior to the implementation of DUFB. In other words, if SNAP customers within the treatment group appear to behave similarly to SNAP customers in the control group prior to the start of the DUFB incentive, then any changes in produce can be more credibly attributed to the incentive and not some unobservable group-specific confounder.
1.2.8 Balance Between DUFB and non-DUFB Stores

Store selection was done prior to receiving any transactions data. Of some concern was that selection stores into the “control” group on a small set of observable variables could still result in drastically different customers, particularly when it came to loyalty card usage rates and spending patterns within the SNAP subpopulation. Some of these concerns were largely covered in the Loyalty Card Concerns section. However, to emphasize the balance between customers visiting DUFB (“treatment”) and non-DUFB (“control”) stores, let’s revisit loyalty card usage rates again but without flattening the data i.e. removing the temporal component. Let’s also do a balance check of average spending per customer transaction.

1.2.9 Loyalty Card Usage Rates

Monthly loyalty card usage rates are calculated by dividing observed loyalty card transactions by total observed transactions. The rate for each store is plotted and a simple loess smoothing curve across time is applied by each group of interest. Stores are split into non-DUFB and DUFB. Stores in the non-DUFB group that were matched via Coarsened Exact Matching or the Linear Probability Model score are represented as circles. On the DUFB group side, circles denote stores that were neither self-selected or unknown. Triangles denote stores that were added retroactively in the non-DUFB group and self-selected/unknown for the DUFB group.

Figure 1.9 shows the over-time average trend between non-DUFB and DUFB stores. The trend between the two stores is relatively similar. The variance in the non-
DUFB group is larger but in both cases, most of the density reside 80 to 85 percent. As noted before, what matters is that loyalty card rates are high across most stores and that high rate persists overtime. In both cases, usage rates remain flat before slightly decreasing in the summer and then slightly increasing towards the end of the calendar year. The reason for this is unclear. I would expect rates to remain flat across time, even despite holiday months like November and December.

It must be emphasized that how the trends co-move is of greater importance than the average value itself. The method of analysis does not require that there be balance in average values. This is not a randomized control trial. It does, however, require balance in how trends move overtime. In this case, the trends are very similar; correlation of the trends is high—0.84.
Similar plots are then generated for the store averages of each observed loyalty transaction over time. The values plotted are averages of averages. Specifically, the average size in dollars of each transactions per loyalty card ID is calculated then these loyalty card averages are average again per store per month. This summarizes average individuals spending behavior up to the store level.

![Graph showing average loyalty transaction sizes in dollars](image)

**Figure 1.10:** Average Number of Observed Loyalty Card Transaction per Customer Averaged per Store (Only SNAP Transactions)

Compared to Figure 1.9, Figure 1.10 subsets to non-SNAP transactions and Figure 1.11 subsets to SNAP transactions only. As with loyalty card usage rates, the co-movement between trend lines is high—0.76 for non-SNAP transactions and 0.94 for SNAP transactions. Correlation between non-DUFB and DUFB SNAP transactions spending sizes is remarkably high and remarkably similar in magnitude. This is a promising result because this subset is the target of interest. It displays that SNAP customers spending is
not dramatically different between non-DUFB and DUFB stores. Of course, if the DUFB incentive is working, then spending on produce (which is not displayed here) should display a break in trend (see Figure 1.12 above for evidence of a break in trend).

Overall, considering the similar trends in loyalty card usage and average loyalty transaction spending sizes over time, there appears to be acceptable balance between the non-DUFB (“control”) and DUFB (“treatment”) stores. Again, balance in the value over time is not a requirement, but the fact that values are relatively similar between the two groups over time helps build confidence that any difference observed in spending on certain produce products at the individual loyalty card level are the result of the DUFB incentive and not some store-level idiosyncrasies.
1.2.10 Produce Spending

The policy goal behind the DUFB incentive program is to increase the produce consumption of SNAP households. A prerequisite to increasing produce consumption is increasing produce spending. To measure whether or not produce spending increased due to the DUFB program, it is important to have a general understanding of produce purchasing patterns across stores by DUFB implementation.

Store-level Produce Spending

![Average of Any Dollars Spent on Produce as a Percentage of All Dollars (2016)](image)

**Figure 1.12:** Average of Store Dollars Spent on Produce as a Percentage of All Dollars.

As shown before, SNAP spending accounts for a small fraction of total dollars spent within any given store. Even assuming a large impact by the DUFB incentive on SNAP produce spending, it should be difficult to observe any change in pooled produce...
spending as a percentage of all dollars spent. There are not enough SNAP dollars being spent to shift the mean of the pooled distribution.

Figure 1.12 plots monthly 2016 produce spending as a percentage of all dollars spent, grouped by DUFB assignment. Dollars are pooled together i.e. no distinction is made between SNAP and non-SNAP dollars. The line represents average produce spending within each group. Of interest is how the gap between each line remains roughly the same. This implies that, throughout 2016, produce spending remains relatively stable, even during the treatment period (which coincides with August to December 2016).

If the intended customers of the DUFB incentive program were any customer purchasing produce, this stable gap between the two lines in Figure 1.12 would be disappointing. It would be evidence, under a difference-in-difference framework, that the DUFB program likely had no impact. But that was not the intended customer group of the DUFB incentive. The intended group was SNAP customers, a small subset of all customers. Therefore, what Figure 1.12 shows is that, if there is a trend shift when comparing produce spending between SNAP customers only, it is likely not the cause of some random confounder that happens to coincide with the DUFB treatment period.

Figure 1.13 replicates Figure 1.12 but subsets to SNAP spending. The lines are the average of the percentage of SNAP dollars spent on produce per month in each store by DUFB assignment. Like Figure 1.12, the gap between the average line in Figure 1.13 is stable from January through July 2016. But unlike Figure 1.12, the gap in Figure 1.13 shrinks between August to December 2016—the DUFB implementation months. The av-
Average of the percentage of SNAP dollars spent on produce within stores assigned to DUFB and during the DUFB intervention is denoted by the dashed green line.

If the DUFB incentive “worked” it is because it increased the amount of dollars spent on produce at the customer level. But an observable collapse in the gap after aggregating SNAP dollars at the store level is evidence that the DUFB incentive may have an impact at the customer level. Store level trends, it can reasonably be assumed, are the result of aggregating customer level trends. The differences-in-differences-in-differences (aka “triple differences”) framework using individual data, after controlling for as many individual customer- and store-level factors, will help determine if the collapsing gap observed in Figure 1.13 is due to DUFB or the result of random coincidental noise. These results are discussed in the Results section.
Customer-level Produce Spending

It is important to understand monthly produce spending at the customer level to provide a reference point before measuring the possible effect of the DUFB incentive on produce spending. Monthly spending per customer per store is used as the dependent variable in the regression analysis. The figures below summarize monthly customer spending on produce and non-produce. The figures therefore help contextualize the magnitude, if any, of the DUFB incentive.

Monthly produce and non-produce spending is calculated (aggregated) per customer within each store then summarized by the median, mean, and/or the transformed zero-modified lognormal mean.\textsuperscript{12} Spending data are heavily right skewed where the strictly positive values are lognormal. Therefore, the zero-modified lognormal mean is used because it better reflects the monthly spending most likely observed per store and helps to damped the impact of outliers in the raw data. This is less of a concern in the regression analysis where variations in customer spending are controlled with fixed effects.

Customer were also grouped by SNAP status prior to store-level aggregation. Customer SNAP status can be either “observed” (“obs”)—meaning observed using SNAP benefits at least once in 2016—or “never”. This is distinct to prior plots that grouped transactions as SNAP or non-SNAP. This approach treated each transaction as independent, ignoring if transactions were purchased by the same customer or household (loyalty ID). But

\textsuperscript{12}For more on lognormal and zero-modified lognormal distributions see Aitchison (1955). Monthly spending by store with zeros representing “nothing spent” have a strong right skew similar to a lognormal distribution. The zero-modified version is a correction to account for the fact that a log normal distribution cannot map to zero.
providing context about produce spending at the customer level necessitates the added step of summarizing spending across customers (loyalty IDs) prior to summarizing at the store level.

For example, if the DUFB incentive effect estimate is found to be 1 dollar more per month, it would matter if average monthly customer spending was 5 dollars versus 20 (a 20% increase versus 5% increase). Likewise, if the units of analysis were in proportion of dollars spent on produce, a 5% increases is rather meaningless if one has no reference point. If the relative effect size is 5% but average monthly spending is only 5 dollars, then the absolute effect size is only 25 cents more per month. Such a change in spending is highly unlikely to lead to any meaningful changes in produce consumption per month.

Each small point in Figure 1.14 represents a transformed lognormal mean of customer monthly produce spending per store by customer SNAP status and store DUFB assignment. The total amount of points per row is 46 (17 DUFB, 29 non-DUFB). The additional point outlined in black is the average of the smaller points cluster within each violin blob. Across all groups, the smaller points fall between 2.5 and 10 dollars. The larger variance of the non-DUFB store group pulls the mean of each cluster slightly above 5 dollars. For the DUFB stores, the mean of each cluster is at or slightly below 5 dollars.

Figure 1.15 follows a similar structure but each small point represents the transformed lognormal mean of customer monthly non-produce spending per store. Almost all small points fall between 20 and 100 dollars. The larger black-stroke point tends to hover between 40 and 45 dollars for all the sub groups except the SNAP observed, non-DUFB
points which hover at or above 50. I do not find this result very surprising. Most of the non-DUFB stores come from more rural areas where there are fewer options to buy non-produce goods. Transaction spending is also higher on average when SNAP participating households spend their SNAP benefits. The interaction of both implies that SNAP participating households in rural areas likely depend and spend more on each store. This drive up total monthly spending per customer, which in turn drives up the store-level mean.

Figure 1.14: Transformed Log-Normal Store Average of Monthly Customer Spending on Produce Grouped by DUFB Assignment, SNAP Status, and Month

Figure 1.16 normalizes customer monthly spending on produce by total monthly spending. As before, each small point represent a store, but the metric is the store-level average of customer monthly spending on produce a percentage of total spending. There is more stability across clusters when produce spending is normalized. On average, SNAP observed households spend less on produce as a percentage of their monthly totals. The
The difference between SNAP observed and corresponding SNAP never observed is roughly 2%. This falls in line with the two prior unnormalized plots, where SNAP produce spending in dollars was lower or the same as non-SNAP produce spending but SNAP non-produce spending generally slightly higher. This would net out to a lower percentage spent on produce for SNAP observed households.

The three plots together help build a reference point to later contextualize regression results in the analysis section. A table with the store average of monthly costumer produce spending, Table 1.7, provides point estimates for reference. The values correspond with the black circles in Figure 1.14. Finally, a table of of non-produce spending, Table 1.8, helps contextualize regression results of non-produce spending, where each value corresponds to the black circles in Figure 1.15.
Figure 1.16: Store Average of Monthly Customer Spending on Produce as a Percentage of Total Spending Grouped By DUF B Assignment, SNAP Status, and Month.

Table 1.7: Average Monthly Produce Spending per Loyalty ID

<table>
<thead>
<tr>
<th></th>
<th>No DUF B</th>
<th>Yes DUF B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No SNAP</td>
<td>Yes SNAP</td>
</tr>
<tr>
<td>Jan 2016</td>
<td>$6.23</td>
<td>$5.90</td>
</tr>
<tr>
<td>Feb 2016</td>
<td>$5.80</td>
<td>$5.61</td>
</tr>
<tr>
<td>Mar 2016</td>
<td>$5.81</td>
<td>$5.75</td>
</tr>
<tr>
<td>Apr 2016</td>
<td>$5.82</td>
<td>$5.86</td>
</tr>
<tr>
<td>May 2016</td>
<td>$6.32</td>
<td>$6.31</td>
</tr>
<tr>
<td>Jun 2016</td>
<td>$6.49</td>
<td>$6.56</td>
</tr>
<tr>
<td>Jul 2016</td>
<td>$6.29</td>
<td>$5.99</td>
</tr>
<tr>
<td>Aug 2016</td>
<td>$5.24</td>
<td>$5.39</td>
</tr>
<tr>
<td>Sep 2016</td>
<td>$4.90</td>
<td>$4.81</td>
</tr>
<tr>
<td>Oct 2016</td>
<td>$5.05</td>
<td>$4.98</td>
</tr>
<tr>
<td>Nov 2016</td>
<td>$5.21</td>
<td>$4.84</td>
</tr>
<tr>
<td>Dec 2016</td>
<td>$5.64</td>
<td>$5.16</td>
</tr>
</tbody>
</table>

Transformed Log-Normal Mean.
Grouped by DUF B Assignment, SNAP Status
### Table 1.8: Average Monthly Non-Produce Spending per Loyalty ID

<table>
<thead>
<tr>
<th></th>
<th>No DUFB</th>
<th></th>
<th>Yes DUFB</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No SNAP</td>
<td>Yes SNAP</td>
<td>No SNAP</td>
<td>Yes SNAP</td>
</tr>
<tr>
<td>Jan 2016</td>
<td>$51.09</td>
<td>$66.35</td>
<td>$40.75</td>
<td>$48.84</td>
</tr>
<tr>
<td>Feb 2016</td>
<td>$48.24</td>
<td>$63.75</td>
<td>$38.53</td>
<td>$46.82</td>
</tr>
<tr>
<td>Mar 2016</td>
<td>$47.26</td>
<td>$61.95</td>
<td>$37.75</td>
<td>$45.87</td>
</tr>
<tr>
<td>Apr 2016</td>
<td>$46.62</td>
<td>$61.50</td>
<td>$38.14</td>
<td>$45.81</td>
</tr>
<tr>
<td>May 2016</td>
<td>$46.55</td>
<td>$59.23</td>
<td>$38.13</td>
<td>$44.98</td>
</tr>
<tr>
<td>Jun 2016</td>
<td>$46.05</td>
<td>$58.20</td>
<td>$37.86</td>
<td>$44.11</td>
</tr>
<tr>
<td>Jul 2016</td>
<td>$46.92</td>
<td>$57.54</td>
<td>$38.24</td>
<td>$43.93</td>
</tr>
<tr>
<td>Aug 2016</td>
<td>$45.76</td>
<td>$58.48</td>
<td>$37.42</td>
<td>$43.86</td>
</tr>
<tr>
<td>Sep 2016</td>
<td>$44.97</td>
<td>$57.03</td>
<td>$37.31</td>
<td>$43.20</td>
</tr>
<tr>
<td>Oct 2016</td>
<td>$45.05</td>
<td>$57.38</td>
<td>$37.75</td>
<td>$43.58</td>
</tr>
<tr>
<td>Nov 2016</td>
<td>$47.29</td>
<td>$59.00</td>
<td>$38.69</td>
<td>$43.96</td>
</tr>
<tr>
<td>Dec 2016</td>
<td>$53.48</td>
<td>$65.47</td>
<td>$42.88</td>
<td>$47.55</td>
</tr>
</tbody>
</table>

Transformed Log-Normal Mean.
Grouped by DUFB Assignment, SNAP Status

### Table 1.9: Average Monthly Proportion of Spending on Produce per Loyalty ID

<table>
<thead>
<tr>
<th></th>
<th>No DUFB</th>
<th></th>
<th>Yes DUFB</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No SNAP</td>
<td>Yes SNAP</td>
<td>No SNAP</td>
<td>Yes SNAP</td>
</tr>
<tr>
<td>Jan 2016</td>
<td>0.10</td>
<td>0.08</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>Feb 2016</td>
<td>0.10</td>
<td>0.08</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>Mar 2016</td>
<td>0.10</td>
<td>0.08</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>Apr 2016</td>
<td>0.10</td>
<td>0.08</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>May 2016</td>
<td>0.11</td>
<td>0.09</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>Jun 2016</td>
<td>0.11</td>
<td>0.09</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>Jul 2016</td>
<td>0.11</td>
<td>0.09</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>Aug 2016</td>
<td>0.10</td>
<td>0.08</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>Sep 2016</td>
<td>0.09</td>
<td>0.07</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>Oct 2016</td>
<td>0.09</td>
<td>0.07</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>Nov 2016</td>
<td>0.09</td>
<td>0.07</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>Dec 2016</td>
<td>0.09</td>
<td>0.07</td>
<td>0.08</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Mean Grouped by DUFB Assignment, SNAP Status
1.3 Empirical Strategy and Methods

The method of analysis requires that the data be at the customer (loyalty card ID) level. The data in its raw form is at the item level where each item is tied to a transaction and, if available, a loyalty ID card number. Such granularity was not necessary. The data was aggregated by loyalty ID card number. Specifically, the data were grouped by loyalty card number, month, and whether the item purchased was produce or non-produce. Other higher level characteristics, like store where the transaction occurred, were also included in the grouping. Spending was then summed up, aggregating out item and transaction level variables.

Spending for each customer was tagged across three groups. The first was whether spending occurred in a store assigned to the DUFB assignment or not. The second dimension was whether spending occurred during the DUFB treatment period (i.e. between the months of August and December 2016). The third dimension was SNAP participation. These three dimension structure the data for use in a difference-in-difference-in-differences (triple differences) analytical framework.

Tagging customers along the first and second groups was straightforward. All store purchases were linked to customers via loyalty ID card numbers. Store assignment to DUFB for each year is known as are the months of the incentive. In fact, in the data, it is possible to identify the first purchase to trigger the DUFB incentive down to the minute. In 2016, the first few DUFB transactions occur in mid July 2016, 2 week prior to when the program was theoretically slated to begin. Since dollars spent were aggregated to the monthly
level for each customer, the DUFB treatment period was tagged beginning in August 2016, leaving the few July 2016 purchases to serve as a sort of “ramp up” to the program being in full swing.

The data are detailed enough that it is not only simple to accurately tag things like store assignment or when the DUFB program is operating, it is also possible to check theoretical assumptions by poking around in the data. Instead of taking on faith that certain stores were assigned to the DUFB program, or that the program ran from August 2016 to December 2016, it is possible to verify in the data directly. Tagging whether a customer was a current SNAP participant, however, was more complicated.

SNAP participation status was inferred from the data, tagging customers observed using SNAP dollars as “SNAP participants” and customers never observed using SNAP dollars as “non-SNAP participants”. SNAP eligibility and participation is, for most US households, temporary. Households can be participants in SNAP anywhere from 3 months to multiple years on a 3 to 6 month renewal, where the majority of households are on a 6 month renewal (Lauffer, 2017). In more practical terms, SNAP participation is actually a dynamic variable, not a static categorical variable. Furthermore, the spending of SNAP dollars serves only as a proxy for a customer’s SNAP participation. It is not possible to observe nor link SNAP dollars being spent by customers in stores outside of the stores for which the data are provided.

As one should expect, there were multiple instances of loyalty cards being observed using SNAP dollars in some months but not others. In other cases, customers were
not observed at all for some months. For example, if 1 denotes a month where SNAP dollars were observed being spent by a customer, 0 denotes a month where spending was observed but never with SNAP dollars, and X denotes no spending observed at all, it is very possible to observe within a span of 12 months a customer with a sequence like X1XX101X0011. The first value in the sequence is January, the last value is December. A customer with sequence X1XX101X0011 is observed for 8 of the 12 months in 2016, 4 during the inactive DUFB months (Jan - July) and 4 during the active DUFB months (Aug - Dec). Of the 8 observed months, 3 were observed not spending any SNAP dollars (three 0) and 5 were observed spending SNAP dollars (five 1).

The problem is that the difference-in-difference-in-differences model framework requires static grouping variables. That is, even though SNAP participation status is dynamic, the framework forces one to impose a strong assumption—that SNAP participation status does not change in the months before and months after the treatment. Given the uncertainty around a customers month-to-month SNAP eligibility and participation, loyalty IDs were tagged as SNAP participants if the loyalty ID was observed using SNAP dollars at least once before the DUFB program began (January - July 2016) and at least once during (August - December 2016). As mention, this does not mean that these customers were always observed using SNAP benefits to make purchases, a prerequisite for earning DUFB points. But that was precisely the goal: the aim was to observe how the DUFB incentive does or does not change increase spending on produce per household overall, not whether spending changed only when SNAP benefits were used.
1.3.1 Store Switching Concerns

The triple differences framework also depends on customers responding to the DUFB incentive without switching stores. The framework would break, for example, if the incentive was so enticing that SNAP customers switched from visiting their usual non-DUFB store to a DUFB store.

Before verifying within the data, store switching was not expected to be a serious problem. The incentive was small and unlikely to overcome the inertia of routine or offset the cost of traveling to a more distant-but-DUFB-eligible store. If such switching did occur, I also expected it would be so rare that corresponding loyalty IDs could be dropped from the data without biasing estimates.

For all loyalty IDs, the set of stores observed in the first half of 2016 were compared to the set of stores observed in second half of 2016. Sets that were unequal were considered possibly problematic. Simply counting how many unique stores were observed was not enough. Counting unique stores would have failed to capture, for example, loyalty IDs that had visited a single non-DUFB stores in the first half of 2016 but then started visiting a single, but different, DUFB store in the second half of 2016. That said, comparing unique DUFB store visits was a quick way of filtering out violating loyalty IDs. Of the roughly 85,000 loyalty IDs observed using SNAP benefits in 2016, about 2.1% were observed switching from visiting zero DUFB stores in pre-DUFB period (Jan - July 2016) to at least one DUFB store during the DUFB period (Aug - Dec 2016). These switching household were dropped.
It is important to keep in mind that the likelihood of observing a loyalty ID visiting 2 or more different stores within a year is incredibly small. Table 1.10 displays a count of loyalty IDs by how many different stores were visited by each in 2016. Roughly 70% of loyalty IDs are observed visiting the same store for all of 2016 and 90% of loyalty IDs are observed visiting at most 2 stores.

Loyalty IDs observed shopping at the same set of 2 or more stores in the first and second half of 2016 were not assumed to violate the triple difference assumptions. For example, it is acceptable for a loyalty ID to be observed visiting only stores \( \{A,B\} \) in the first half of 2016 and then again only stores \( \{A,B\} \) in the second half.\(^{13}\) Also note that the triple-difference model requires there be at least one observation per loyalty ID and store ID pair in both the pre-DUFB and DUFB periods. To help with identification, I dropped any \([\text{loyalty ID}, \text{Store ID}]\) pairs that were not observed at least twice in both the pre-DUFB and DUFB periods. I also dropped loyalty IDs observed visiting more than 4 different stores within the year. This results in a loss of 461,768 loyalty IDs which accounts for 55% of all loyalty IDs and 37% of raw monthly panel data.\(^{14}\)

**Table 1.10**: Count of Loyalty IDs by Stores Visited in 2016

<table>
<thead>
<tr>
<th>visited</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.689</td>
</tr>
<tr>
<td>2</td>
<td>0.206</td>
</tr>
<tr>
<td>3</td>
<td>0.072</td>
</tr>
<tr>
<td>4</td>
<td>0.023</td>
</tr>
<tr>
<td>5+</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\(^{13}\)Operationally, I generated the set difference, `setdiff()`, between the store IDs observed in the first and second halves of 2016 for every loyalty ID.

\(^{14}\)This may seem like too large a trim, but the vast majority of observations dropped are those that would have been dropped anyway; Stata would have dropped these automatically as “singleton” observations.
1.3.2 Variables and Data Structure

The specific structure of the data is best explained by displaying a few rows of the data and substituting sensitive variables with fake, but realistic, values. The variables \texttt{loyalty\_id}, \texttt{year}, and \texttt{month} are included in Table 1.12 to help provide context.

The variable \texttt{year} has no impact on any model; \texttt{year} does not vary in this model. Variables \texttt{loyalty\_id} are used to generate individual-level fixed effects (\texttt{fe\_i}) and \texttt{month} is used in constructing month fixed-effects (\texttt{fe\_t}). All the dummy variables (prefixed by \texttt{d\_}) are used to construct the model framework, including variables \texttt{ddd} and \texttt{ddd\_t}, which help measure the effect of the DUFB program.

\begin{table}[h]
\centering
\begin{tabular}{cccccccc}
\hline
\texttt{loyalty\_id} & \texttt{year} & \texttt{month} & \texttt{y\_ijt} & \texttt{fe\_i} & \texttt{fe\_j} & \texttt{fe\_t} & \texttt{ddd\_t} & \texttt{dd\_ij} & \texttt{dd\_jt} & \texttt{dd\_it} \\
\hline
87XXXXXX & 2016 & 11 & 15.49 & 87XXXXXX & 1586 & 11 & 1 & 1 & 1 & 1 \\
87XXXXXX & 2016 & 11 & 0 & 87XXXXXX & 1586 & 11 & 11 & 1 & 1 & 1 \\
43XXXXXX & 2016 & 9 & 78.78 & 43XXXXXX & 1523 & 9 & 9 & 1 & 1 & 1 \\
43XXXXXX & 2016 & 9 & 0 & 83XXXXXX & 1523 & 9 & 9 & 1 & 1 & 1 \\
98XXXXXX & 2016 & 12 & 64.24 & 98XXXXXX & 1990 & 12 & 12 & 1 & 1 & 1 \\
98XXXXXX & 2016 & 12 & 0 & 98XXXXXX & 1990 & 12 & 12 & 1 & 1 & 1 \\
52XXXXXX & 2016 & 12 & 45.8 & 52XXXXXX & 1979 & 12 & 0 & 0 & 0 & 1 \\
52XXXXXX & 2016 & 12 & 0.28 & 52XXXXXX & 1979 & 12 & 0 & 0 & 0 & 1 \\
14XXXXXX & 2016 & 12 & 171.19 & 14XXXXXX & 1524 & 12 & 0 & 0 & 1 & 0 \\
14XXXXXX & 2016 & 12 & 0 & 14XXXXXX & 1524 & 12 & 0 & 0 & 1 & 0 \\
96XXXXXX & 2016 & 12 & 58.6 & 96XXXXXX & 123 & 12 & 0 & 0 & 0 & 1 \\
96XXXXXX & 2016 & 12 & 8.98 & 96XXXXXX & 123 & 12 & 0 & 0 & 0 & 1 \\
\hline
\end{tabular}
\caption{Sample of Rows from Data Set (Columns 1 - 11)}
\end{table}

Recall that of interest is the impact the DUFB incentive program had on customer (loyalty ID) \textit{produce} spending. For every observed loyalty card transaction, monthly spending is split into two categories, produce vs non-produce. The sample table above, Table 1.12, includes produce (odd rows; not highlighted) and non-produce (even rows; highlighted) spending. In the table, spending is the dependent variable labeled \texttt{y\_ijt}. If spending is observed for a loyalty card ID but the monthly sum of either category is missing, then a
zero is imputed. The decision not to spend money on a given category within a month by a
given customer is therefore included in the model.

The variable names in Table 1.12 show the three dimensions the data can possibly vary: by month, by customer, and/or by store. The dependent variable $y_{ijt}$ (spending) varies along all three. The independent variables, however, can vary by one, two, or three. The independent variables are tagged according to the dimensions they vary along. Variables that vary month to month contain a $t$ in their names; variables that vary across customers (loyalty card IDs) contain an $i$; and variables that vary by store contain a $j$. Variables that vary along all three dimensions contain $ijt$ in their names e.g. $y_{ijt}$. Others vary by only two dimensions and contain, for example, $jt$, denoting variation by store and by month, or $it$, denoting variation by loyalty ID and by month. The implications and reasoning behind labeling these variables will be made clear in the Methods section.
1.3.3 Methods

Difference-in-difference-in-differences (DDD or triple-difference) is the model framework used to measure the effect of the DUFB incentive program on produce purchases (Wooldridge and Imbens, 2007). Two flavors of the DDD model were used. The first, and more traditional, splits time into two periods: before and after the start of the “intervention”. In this case, the two periods would be before the DUFB program was implemented (pre-August 2016) and the period after the DUFB program was in effect (August 2016 and onwards).

The second flavor splits the intervention period into its individual months. The pre-intervention (pre-August 2016) period remains the same but the intervention period, in this specific case, is split into 5 different month dummy variables (August through December 2016). This design helps measure if/how the impact of the DUFB incentive possibly changes over time. That is, it helps show if the DUFB incentive has a short-run effect that changes over time or if the effect persists at constant value. As hypothesized earlier, the effect was expected to fade over time.

**Difference-in-Difference-in-Differences (DDD) Model 1**

Let $y_{ijt}$ be total monthly produce for customer $i$ in store $j$ during month $t$. For every customer $i = 1, \ldots, N$ in store $j = 1, \ldots, J$ across $t = 1, \ldots, T$ months. The DDD regression is
\[ y_{ijt} = \alpha_i + \gamma_j + \lambda_t + \rho_1 dS_i + \rho_2 dU_j + \rho_3 dP_t \\
+ \theta_1 dS_i \cdot dU_j + \theta_2 dP_t \cdot dU_j + \theta_3 dP_t \cdot dS_i \\
+ \delta dU_j \cdot dS_i \cdot dP_t + z'_{jt} \phi + \epsilon_{ijt} \]

where \( dS_i \) represents SNAP/EBT participant status (target group), \( dU_j \) represents store assignment to DUFB group, and \( dP_t \) represent the treatment period, August - December. \( \alpha_i \) captures customer (individual) fixed-effects, \( \gamma_j \) captures store fixed-effects, and \( \lambda_t \) captures monthly (time) fixed-effects. \( z'_{jt} \) is a vector of observable store-level controls for store \( j \) during month \( t \). The coefficient of interest is \( \delta \). \( \epsilon_{ijt} \) are idiosyncratic errors at the customer-level.

By design, the various fixed effects capture variables \( dU_j, dS_i, \) and \( dP_t \). To simplify further, compress \( dS_i \cdot dU_j \) to \( D_{ij} \). Likewise compress the remaining dummy variables to \( D_{jt}, D_{it}, D_{ijt} \), representing the dimensions along which they vary—\( i \) customer loyalty ID, \( j \) store DUFB assignment group, and \( t \) DUFB treatment period (before or during August - December 2016). The model becomes

\[ y_{ijt} = \alpha_i + \gamma_j + \lambda_t \\
+ \theta_1 D_{ij} + \theta_2 D_{jt} + \theta_3 D_{it} \\
+ \delta D_{ijt} + z'_{jt} \phi + \epsilon_{ijt} \]

**Difference-in-Difference-in-Differences (DDD) Model 2**

To capture more detail than just the average, \( \delta_t \) is allowed to vary by each month that the DUFB program is active. The model (first term, third row of the equation) changes slightly
The interpretation of $t$ for each of the dummy variables $D_{jt}$, $D_{st}$, $D_{ijt}$ changes. Instead of denoting the full span of months either before or during the DUFB treatment period, it separately denotes each month of the DUFB program, similar to the monthly fixed-effect variable. The purpose of this second DDD model is to estimate the effect of the incentive over time. The first model averages the effect across the entire treatment period, August - December 2016. The second model separates out the effect of each month, differentiating between the average effects in August 2016 versus, for example, November 2016. It is reasonable to assume that the effect of the program was not constant over time and it is useful to understand how the effects of the program (likely) fade in the long term.

### 1.4 Results

The 2016 data contained 3.39 million rows of both produce and non-produce sales after being aggregated to the loyalty-month-store level. Two regression tables are produced below, both estimate using the `reghdfe` Stata package (Correia, 2017).\(^{15}\) Tables were generated for both produce and non-produce regression results.

All models, as described in the [methods][#methods] section, contain individual

\(^{15}\)Estimates were also generated using the R package `lfe` as a sanity and consistency check (Gaure, 2013). The tables were excluded because they had identical results.
Table 1.13: Produce Spending Triple Difference Estimates

<table>
<thead>
<tr>
<th></th>
<th>Model 1 (ddd)</th>
<th>Model 2 (ddd_t)</th>
<th>Model 3 (ddd)</th>
<th>Model 4 (ddd_t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ddd )</td>
<td>0.334**</td>
<td>0.305**</td>
<td>0.535***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.15)</td>
<td>(0.17)</td>
<td></td>
</tr>
<tr>
<td>( ddd_{t08} )</td>
<td>0.504**</td>
<td>0.535***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( ddd_{t09} )</td>
<td>0.633***</td>
<td>0.561***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( ddd_{t10} )</td>
<td>0.494***</td>
<td>0.366**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( ddd_{t11} )</td>
<td>0.341*</td>
<td>0.267</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.18)</td>
<td>(0.18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( ddd_{t12} )</td>
<td>-0.343***</td>
<td>-0.239*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_{jt \log ebt _sales _max} )</td>
<td>-0.339***</td>
<td>-0.333***</td>
<td>0.351***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.15)</td>
<td></td>
</tr>
<tr>
<td>( x_{jt \log n _loyalty _ids} )</td>
<td>-4.925***</td>
<td>-4.796***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.41)</td>
<td>(1.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_{jt \log sales _ebt _produce} )</td>
<td>1.321**</td>
<td>1.304**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.60)</td>
<td>(0.60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_{jt \log sales _produce} )</td>
<td>8.733***</td>
<td>8.663***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.25)</td>
<td>(1.25)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* \( p < 0.1, ** p < 0.05, *** p < 0.01 \)

‘ddd’ and ‘ddd_t’ estimates. Generated with Stata package ‘reghdfe’.

(loyalty ID), store, and month fixed effects. The first two columns of each table are regression models without any additional regressors. The last two columns of each table add time-varying store-level regressors. Although estimated, no models with time-varying individual-level regressors were included in the tables. All models used cluster-robust standard errors, where clustering was done at the store level.

It was impossible to generate any time-varying individual-level regressors that were not, directly or indirectly, affected by the DUFB incentive. Models with these individual controls were excluded from the final results of this paper. Better to totally avoid the sin of “conditioning on the treatment” (aka “over-control bias”) than to go stargazing through poorly constructed models (Elwert and Winship, 2014).
### Table 1.14: Non-Produce Spending Triple Difference Estimates

<table>
<thead>
<tr>
<th></th>
<th>Model 5 (ddd)</th>
<th>Model 6 (ddd_t)</th>
<th>Model 7 (ddd)</th>
<th>Model 8 (ddd_t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ddd</td>
<td>1.039</td>
<td>1.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.79)</td>
<td>(0.75)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ddd_t08</td>
<td>2.845**</td>
<td>3.309***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.21)</td>
<td>(0.97)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ddd_t09</td>
<td>2.814***</td>
<td>2.893***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.84)</td>
<td>(0.81)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ddd_t10</td>
<td>1.653*</td>
<td>1.217</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.91)</td>
<td>(0.89)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ddd_t11</td>
<td>1.520</td>
<td>1.131</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.92)</td>
<td>(0.92)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ddd_t12</td>
<td>-3.939***</td>
<td>-3.816***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.00)</td>
<td>(0.98)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x_jt_log_ebt_sales_max</td>
<td>-1.009</td>
<td>-0.961</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.67)</td>
<td>(0.66)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x_jt_log_n_loyalty_ids</td>
<td>-5.920</td>
<td>-4.829</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5.50)</td>
<td>(5.46)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x_jt_log_sales_ebt_produce</td>
<td>1.475</td>
<td>1.340</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.21)</td>
<td>(2.19)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x_jt_log_sales_produce</td>
<td>33.299***</td>
<td>32.702***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5.12)</td>
<td>(5.08)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p < 0.1, ** p < 0.05, *** p < 0.01

‘ddd’ and ‘ddd_t’ estimates. Generated with Stata package ‘reghdfe’.

Monthly produce spending possibly increased by about 31 cents per loyalty ID, on average, during the treatment period (Aug - Dec 2016). The “Model 3” column of Table 1.13 contains estimates for the treatment effect averaged across all active months, denoted ddd. Neither estimate of the ddd coefficients are significant.

The section on produce spending provides context for these values. The average SNAP customer visiting the average DUBF store spent about 4 dollars per month on produce from Jan - July 2016. An increase of 31 cents would therefore corresponds to about an 7.75% increase in total monthly produce spending.

Measuring the effect of the incentive in dollars versus percentages is important. Percentages alone mask how small the effect size is when translated to purchasing power. 31
cents is a very small amount. Checking prices on-line from the participating grocery store, 31 cents would correspond to, for example, about 1 more bananas purchased per month or about 1 oz of green beans.

The average effect size over the entire treatment period, $ddd$, however, masks how the effect size fades over time. The “Model 4” column Table 1.13 contains estimates for each month the incentive is active in 2016, denoted $ddd_{t08} - ddd_{t12}$ (the numbers correspond to the months of the treatment, Aug - Dec 2016). Estimates start close to 54 cents in August 2016 before decaying to about 36 cents in October and finally negative 24 cents by December 2016. By November and December, the effect is insignificant, assumed to be no different from zero. Again, in context to what is historically spent on produce per month—$4 on average—the effect overtime peaks at a roughly 13.5% increase in produce spending in August 2016 before decaying to essentially zero by December 2016 (compared to the pre-DUFB treatment period). The DUFB treatment effect therefore appears to have a short-run effect. This follows the pattern observed in other incentive programs that aim to change behavior, like diet/weight-loss and energy conservation interventions (Allcott and Rogers, 2014; Harding and Hsiaw, 2014; Verplanken and Wood, 2006). The short-run (fading) effect stands in contrast to what classical economic theory would suggest: a smooth, constant effect observed across all months of the DUFB program. Instead, the fading effect falls more in line with what is predicted by behavioral economic theory: complexity is a cost that consumers will avoid at the expense of a possible (but not very salient) financial gain.
It is insightful to compare changes in produce spending to non-produce spending under the same framework. Following the same structure as produce regression table, the “Model 7” ddd coefficient estimate of Table 1.14 was roughly 1.04 dollars. The standard errors, however, display a substantially higher degree of uncertainty than those from the produce regression tables (Model 3). Though neither of the ddd coefficient estimates for produce and non-produce are statistically significant, the standard errors capture how much more unlikely it would be, on average, to find any notable increase in non-produce spending.

As with the the produce models, the ddd estimate masks how differences in spending decay from month to month. In the first month of the treatment (ddd_t08) non-produce spending increased by 3.31 dollars. The effect decays to being virtually zero (ddd_t10 and ddd_t11) before going negative in December (ddd_t12). Even at the short-run peak in August 2016, the relative increase in non-produce spending, compared to historical non-produce average—roughly $45—is around 7.36%.

In absolute terms, it is possible for spending to increase across all categories; one expects spending on produce to correlate highly with non-produce spending. The DUFB incentive, if working, likely affects spending at DUFB participating stores in two ways. First, it may incentivize SNAP households to spend more of their total budget at DUFB participating stores. This would take advantage of the incentive during the active months, resulting in increased spending across both produce and non-produce items. Second, if spending does increase across both categories, the expectation, assuming the incentive is working, is that the marginal increase in produce spending, relative to the marginal increase
in non-produce spending, would be higher.

Put another way, assume the average SNAP household spends 50 dollars total a month, $5 on produce and $45 on non-produce (10% vs 90% of average spending). The existence of the DUFB incentive may increase total spending by an extra $4.50 more per month—a total marginal spending increase of 4.5 dollars. However, if the fraction of those $4.50 spent towards produce switched to $1 produce and $3.50 non-produce, the marginal increase relative to the pre-incentive category averages would be $1/5 = .2 (20%) produce and $3.5/45 = .077 (7.7%) non-produce. Likewise, the spending proportions shift from $5/50 = .1 (10%) being spent on produce, on average, to $1/4.5 = .22 (22%) on the margin. Estimates, therefore, of how spending change in absolute terms for both produce and non-produce helps to better explain how the incentive may be affecting spending, particularly substitutions at the margin. With respect to this analysis, the proportion spent on produce on the margin is notably different to the pre-DUFB average. In the pre-DUFB period the proportion of dollars spent on produce was on average around 9.00%. In August 2016, the proportion spent on produce on the margin was \( \frac{.54}{.54 + 3.31} \times 100 = 14.0\% \), increasing to \( \frac{.56}{.56 + 2.89} = 16.2\% \) in September 2016.\(^{16}\)

Overall, there is evidence that the DUFB incentive increased produce spending during the second half of 2016. There is strong evidence that the DUFB incentive has a short-run effect on produce spending given the effect decays over-time. The implications of a relatively small absolute effect size but promising relative effect size, as well as suggested

\(^{16}\)Note that changes in the proportion spent on produce is another way to have structured this model. However, as mentioned, this masks the absolute effect (dollars), which better communicates the potential impact of the DUFB incentive.
improvements to the DUFB program, are discussed in the concluding section.

1.4.1 Heterogeneous Effects

The prior triple-difference models estimate the effect of the DUFB by pooling all shoppers into a single group. This is an appropriate measurement of how a SNAP shopper would, with no a priori information about the shopper, respond to the DUFB incentive if it became available. But to assume we know nothing about shoppers spending habits on produce before the start of the DUFB program is naive. Loyalty ID cards make it possible to segment shoppers by how much they spend in total, and on produce, very easily. For example, it is possible to know how much a specific SNAP shopper was spending on produce before the DUFB incentive became available.

Any major grocery retailer that uses loyalty ID cards would have access to, at a minimum, a shoppers total spending. With these data, it is possible to simulate how any profit-seeking retailer would behave if given access to spending data. A reasonable action to expect of any retailer would be to segment shoppers. One of the easiest segmentations is by total spending. Total spending acts as a literal measure of store loyalty. Shoppers that spend more money tend to visit stores more often. It can also be a reasonable proxy for whether a store is a shopper’s priority grocery destination.

It is reasonable to assume that different SNAP customer segments would respond differently to the DUFB incentive. SNAP households that do their main shopping trips every month at DUFB stores would inherently have more opportunities to learn about, and make use of, the DUFB incentive. We would anticipate these SNAP households to spend
more per month, on average, than SNAP households that spend less of their income and benefits at DUFB stores. In short, it is reasonable to expect that, if segmented into different groups by total spending, we would measure heterogeneous effects of the DUFB incentive across each of the segments.

Segmentation was done via K-means clustering, an unsupervised learning algorithm that clusters a matrix of inputs into different cluster (Hartigan and Wong, 1979). Three clusters were selected. The input matrix was total spending, produce spending, and number of total transactions observed from January to July 2016 (the pre-DUFB period) per loyalty ID. Figure 1.17 shows the segmentation. Groups were split most sharply by average Total Monthly Spending across the 7 months observed before the start of DUFB program. This logically correlates highly with produce spending and total transactions observed. All spending was transformed into log scale to make it easier to plot and compare spending groups.
Figure 1.17: percentage of total dollars spent on sugar-sweetened beverages per month among SNAP recipients shopping in intervention and control stores, 2016
Table 1.15 shows the results of running the same triple-difference model as before but across 3 different groups of shoppers segmented by total spending. Group 1 is the low spending group. It accounts for the vast majority of shoppers that patron the participating grocery chain. Group 2 is the medium spending group and Group 3 is the high spending group. Group 3 contains the fewest shoppers and accounts for the shoppers that the retailer would label as its most loyal segment.

Averaged across the entire DUFB period (Aug - Dec 2016), neither Group 1 or Group 2 display any notable effect of the incentive. Results are under columns 1 and 3 (Group 1 (ddd) and Group 2 (ddd)) labeled row ddd. While generally statistically insignificant, the monthly estimates in columns 2 and 4 (Group 1 (dddt) and Group 2 (dddt)) show the same trend as the pooled estimates: the magnitude of the effect is largest in the first three months of the incentive (Aug - Oct) and then decreases to being negative by December. The regression estimates for Group 3 in columns 5 and 6 are similar in trend but the magnitudes are much higher and the monthly estimates are statistically significant. Proportionally, the effect sizes of each group, relative to the log-normal produce spending average, max out at roughly 5%. That is, the increase in produce spending per group due to the DUFB incentive is at best 5%. The highest spending group (Group 3) increased their produce spending more, in absolute terms, than the the lowest spending group (Group 1) by roughly 10 fold—roughly $0.18$ dollars versus $1.70$ dollars during the second month of the DUFB program (Sept 2016).
Table 1.15: Heterogeneous Produce Spending Triple Difference Estimates

<table>
<thead>
<tr>
<th></th>
<th>Group 1 (ddd)</th>
<th>Group 1 (ddd)</th>
<th>Group 2 (ddd)</th>
<th>Group 2 (ddd)</th>
<th>Group 3 (ddd)</th>
<th>Group 3 (ddd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ddd</td>
<td>0.045</td>
<td>0.041</td>
<td>0.757</td>
<td>0.66</td>
<td>1.695**</td>
<td>1.704**</td>
</tr>
<tr>
<td>ddd_t08</td>
<td>0.165</td>
<td>0.387</td>
<td>0.088</td>
<td>0.29</td>
<td>0.351</td>
<td></td>
</tr>
<tr>
<td>ddd_t09</td>
<td>0.181*</td>
<td>0.484</td>
<td>0.29</td>
<td>0.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ddd_t10</td>
<td>0.177***</td>
<td>0.088</td>
<td>0.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ddd_t11</td>
<td>-0.027</td>
<td>-0.066</td>
<td>0.79</td>
<td>1.255*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ddd_t12</td>
<td>-0.293**</td>
<td>-0.732***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x_jt_log_ebt_sales_max</td>
<td>-0.181***</td>
<td>-0.177***</td>
<td>-0.525***</td>
<td>-1.473***</td>
<td>-1.225**</td>
<td>-1.125**</td>
</tr>
<tr>
<td>x_jt_log_n_loyalty_ids</td>
<td>-2.659***</td>
<td>-2.575***</td>
<td>-6.844***</td>
<td>-11.225**</td>
<td>-10.887**</td>
<td></td>
</tr>
<tr>
<td>x_jt_log_sales_ebt_produce</td>
<td>0.602***</td>
<td>0.596***</td>
<td>1.856**</td>
<td>4.579*</td>
<td>4.504*</td>
<td></td>
</tr>
<tr>
<td>x_jt_log_sales_produce</td>
<td>5.370***</td>
<td>5.319***</td>
<td>11.929***</td>
<td>18.169***</td>
<td>18.004***</td>
<td></td>
</tr>
</tbody>
</table>

*p < 0.1, **p < 0.05, ***p < 0.01

Estimates ‘ddd’ and ‘ddd_t’ generated with Stata package ‘reghdfe’.
1.5 Conclusion

The results of the DUFB incentive program are modest but promising. The effect of the incentive on produce spending was, on average, less than a dollar more per month. In relative (percentage) terms, the effect was similar to what was found in the Healthy Incentives Pilot (HIP)—the seminal study to have investigated whether or not a targeted financial incentive increased spending on fruits and vegetables. HIP found the relative effect to be 11% using the point-of-sale (transaction) data and 8.5% using the survey recall data.

It is not clear what the results imply from a policy perspective. If the DUFB program were to expand to more retail grocery chains, it is reasonable to anticipate an increase in produce spending. But could a 30 cent to 1 dollar per month increase in produce spending result in any notable health improvements in the short- or long-term? That is a more difficult question. For the DUFB program to have notable impact, it would have to be available to SNAP households in all grocery stores, not just within participating retail chains. The relative (percentage) effect on produce spending corresponds with a worthwhile absolute impact only if the incentive is accessible to enough retail environments where SNAP households are likeliest to spend their benefits. An 8% increase on $50 per month spent could make a difference over time, but not an 8% increase on $5. As observed in the results section, the SNAP households that benefited most were the households that spent more of their SNAP EBT benefits at grocery stores participating in the DUFB program. Considering it was not possible to observe where else SNAP households may have been also been spending money, one can only imagine that if the DUFB program were available at every
household’s default grocery store—where a household spends most of its time and money shopping for groceries—a similar increase in produce would be observed. In other words, if the DUFB incentive program was standard and ubiquitous, then SNAP households could be expected to spend at least a dollar more per month.

But what does it mean, from a policy perspective, to increase produce spending by a dollar more per month? Spending itself is a proxy for the actual consumption (ingestion) of food. All food items purchased are at best totally consumed and at worst thrown away. Spending, in other words, leads to a non-negative change in food consumption. Assuming households consume at least 3/4 of the food they purchase, then what would be the value of a households consuming a bundle of bananas more per month? The health and nutrition literature tends to frame things in how many more servings of produce are consumed per day. For example, eating more fruits and vegetables per day is linked to reductions in cancer, heart disease, and stroke (Slavin and Lloyd, 2012; Van duyn and Pivonka, 2000). How does one think about the public health impacts of an incentive that, to be estimated accurately, is measured per month? How does one then translate the nutritional effects of spending one more dollar more per month into food consumed (ingested) per day? These are difficult, if not impossible, conversions to make and measure.

However, despite the modest absolute increase in monthly produce spending, the DUFB incentive program should not be written off. With a few improvements—many of which were adopted by the grocery chain in 2017—the DUFB incentive program could further increase produce spending, and hopefully consumption. More interesting still would
be if the persistent availability of the DUFB incentive, by making fresh produce cheaper, could lead to a shift in household food preferences.

The greatest impediment to the DUFB incentive program being effective isn’t the mechanism or implementation—though it can certainly improve (more on that next)—it is the general lack of demand for produce. Produce just isn’t very popular. Household monthly spending on produce (fruits and vegetables) by SNAP households is, at its best, about 14%. This is according to Garasky et al. (2016), the only other researcher I could find that analyzed point-of-sale data. In that paper, however, Garasky et al. (2016) include fruit juice, fruit snacks, and frozen vegetables as in the “fruit” and “vegetable” categories. This overstates the popularity of fruits and vegetables. This is using the grocery retailer’s definition of produce, which excludes juices and frozen goods. Given how popular fruit juices, fruit snacks, and frozen vegetables are, it could be argued that the produce (fruit and vegetable) spending estimates in this dissertation, which hovers between 5% and 10%, are more accurate and realistic.17

It is possible that the DUFB incentive, by making produce less expensive, increases the likelihood that SNAP households try to incorporate more fresh fruits and vegetables into their diets. In the long run, this may increase the appetite (demand) for these goods, leading to increased produce spending. But for this to feasibly happen, the DUFB mechanism needs to be simplified and expanded. The 2016 mechanism studied in this paper followed the spirit of the original DUFB mechanism first introduced in farmer’s markets.

17The point-of-sale spending estimates on fruits and vegetables in the Healthy Incentives Pilot were also around 5%. These also excluded fruit juices and frozen vegetables. The survey recall estimates, however, were considerably higher at 21% (Bartlett et al., 2014).
At farmer’s markets, SNAP households could “Double Up” their SNAP benefits and then redeem those tokens (points) on locally grown produce. The 2016 grocery store mechanism, like the farmer’s market version, had the “helping local farmers by buying local produce” spirit. It differs in that, while SNAP households must buy local, they can redeem any produce, local or not, in the grocery store version. This “buy local” spirit of DUFB was perhaps the main reasons the DUFB incentive was so politically popular. The mechanism design was considered a win-win for local farmers and local low-income households, and helped DUFB gain FINI pilot program funding.

But the farmer’s market DUFB shopping experience is very different from the retail grocery store experience. First, the DUFB payment mechanism in the farmer’s market is far more salient than the grocery store payment mechanism. In the farmer’s market, recall that one has to exchange SNAP dollars for physical tokens. The process of spending results in a physical reduction in the amount of tokens following every DUFB redemption. This is far more salient than earning and redeeming points in a loyalty card transaction. Second, participation in the farmer’s market DUFB program is narrow and self-selective: SNAP households literally opt-in to participating by buying tokens. Lastly, most, if not all, of the produce is local in the farmer’s market. There is no confusion as to whether one is buying local or not, like there possibly is at the grocery store. In short, while the spirit of the grocery store DUFB program followed the spirit of the original farmer’s market version (buy local), the payment mechanisms differed in ways that made the grocery store DUFB mechanism less salient and less obvious to participants. But things improved in 2017.
In 2017, the grocery stores’ DUFB mechanism changed in two important ways. The first was that shoppers could now earn and redeem on any fresh produce. This made the earning and redeeming mechanism much easier for shoppers to understand. Second, the set of items eligible for earning and redeeming points were much more logically and obviously defined. Before 2017, it was not obvious what was or was not local produce. This obfuscated how SNAP household earned points. But what was considered produce was also not obvious. For example, candied nuts and fruits were considered produced, but not certain kinds of fresh, pre-packaged fruits. This obfuscated which items were qualified for redeeming points. I happen to notice that many strange items were considered produce by the grocery chain. They did not seem aware of these incorrect categorization. It was too late to change for the 2016 data but my digging through the data helped inform how the company identified eligible (fresh) produce. Eligible items are now much more logically and obviously in line with what shoppers would imagine to be healthy fruits and vegetables. Together, these changes to the mechanism reduced the complexity of the incentive program and increased opportunities to learn about the program.

The 2017 change lead to a greater than 10 fold increase in the earning and redeeming of points. This was notable because the participating grocery retailer used the total spending of points as a measure of the program’s popularity and success. Points earned and redeemed, from the grocery chains perspective, were disappointingly low in 2016 and the 2017 change made them more optimistic about the DUFB programs success. Granted, the 2017 mechanism change also coincided with the program being expanded to virtually all
stores. It is hard to disentangle how much of the increase was due to the reduced complexity versus the increased opportunity for SNAP participants to passively (or actively) earn points. I leave it to future researchers to investigate whether or not changing the earning and redeeming mechanism increased total fresh produce sales.
Chapter 2

Evaluating a USDA Pilot Program to Incentivize the Purchase of Fresh Produce Among SNAP Beneficiaries in Supermarkets

2.1 Introduction

Obesity is a strong risk factor for a variety of health conditions, including cardiovascular disease, diabetes, and various cancers (Burke GL et al., 2008; Peytremann Bridevaux and Santos-Eggimann, 2008; Peeters et al., 2003) Though increases have leveled off, the prevalence of obesity remains high in U.S. adults and children (Ogden et al., 2014). Nationally, obesity prevalence is higher among non-Hispanic black and Hispanic adults and youth compared to non-Hispanic white adults and youth; and recurrently poor children are at greater risk for an adverse BMI growth trajectory (Ogden et al., 2014, 2018; Min et al., 2018). Previous studies also suggest that participants in the Supplemental Nutrition Assistance Program (SNAP) are more likely to be obese than non-participants, and children in SNAP-recipient households are more likely to have higher adult BMI values than other low-income children (Nguyen et al., 2015; Vartanian and Houser, 2012).

Diets rich in energy and low in fiber, fruits, and vegetables are positively associated with risk of excess adiposity and higher BMI (Rouhani et al., 2016). However, low-income adults consume fewer fruits and vegetables than higher income adults (Hiza et al.,
2013; Storey and Anderson, 2014). Previous work also indicates that adult SNAP participants purchase fewer fruits, vegetables, and whole grains than higher income households; and purchase more total calories, sodium, and sugars (Mancino et al., 2018; Grummon and Taillie, 2017). Disparities in diet behaviors stem in part from the relatively low price of nutrient-poor, energy-dense foods and beverages and high consumption of food-away-from-home (Finkelstein et al., 2005; Blumenthal et al., 2014; Lin et al., 2014; Racine et al., 2013). Further, fresh produce is less available in poorer communities and, notably, more expensive (Blumenthal et al., 2014; Lin et al., 2014; Racine et al., 2013). Therefore, interventions to increase the affordability of produce may help mitigate disparities in diet and obesity risk, especially among SNAP participants.

To date, only a few studies have tested pricing incentives aimed at promoting healthy food purchases among low-income individuals (Afshin et al., 2017). A prominent example is the Healthy Incentives Pilot, which introduced a 30-cent subsidy for every SNAP dollar spent on eligible fruits and vegetables (Olsho et al., 2016). The evaluation of the pilot study demonstrated that intake of fruits and vegetables increased in SNAP households, and results were cited as support for the development of the Food Insecurity Nutrition Incentive (FINI) program by the United States Department of Agriculture (USDA) (Food and Nutrition Services, 2015). The FINI program supports local efforts that help SNAP participants increase their purchase of produce through financial incentives, including Fair Food Network’s Double Up Food Bucks (DUFB) program in Michigan. Fair Food Network’s DUFB program was first piloted in five Detroit farmers’ markets in 2009 and provided con-
sumers with a one-to-one matching credit up to $20 per market day toward Michigan-grown 
produce for each dollar of benefits spent on SNAP-eligible items. Subsequently, the pro-
gram expanded to more than 150 farmers’ markets in Michigan and more than 20 additional 
states.

In 2015, Fair Food Network was awarded more than $5 million from the USDA
(along with non-federal matching funds) to expand the DUFB program to additional loca-
tions, including supermarkets, over a four-year period (Network, 2014). In prior research,
the DUFB program has been shown to increase self-reported vegetable intake and reduce 
food insecurity among patrons of farmers’ markets (Young et al., 2013; Savoie-Roskos et al.,
2016). However, farmers’ markets are relatively scarce compared to other food outlets and
U.S. adults purchase more than three-quarters of total energy intake from supermarkets
(Byker et al., 2012; Singleton et al., 2015; Drewnowski and Rehm, 2013). In addition, the
majority of farmers’ markets are seasonal (Byker et al., 2012). Therefore, testing produce
incentives for SNAP recipients in supermarkets is a logical next step in evaluating produce
incentives. To our knowledge, only one study has directly tested such an intervention, but
it was restricted to one supermarket in a rural community with no control group and did not
capture changes in less healthy purchases (Polacsek et al., 2018).

To address these gaps in the literature, we sought to examine the impact of Fair
Food Network’s DUFB program on produce purchases among SNAP recipients shopping
in over 30 supermarkets store in the Midwestern United States between 2015 and 2016. We
also examined the impact of the DUFB program on changes in sugar-sweetened beverage

90
(SSB) purchases. We hypothesized that the percent of dollars spent on fresh produce would increase in intervention stores among SNAP participants as a result of the DUFB program.

2.2 Methods

The transaction data used in our study come from 32 supermarkets. All stores come from the same supermarket chain of a major grocery retailer in the United States. In 2014, Fair Food Network piloted the DUFB program in two supermarkets in this chain. In 2015, three supermarkets that were geographically proximal to the two pilot supermarkets were added to the program, for a total of five intervention stores. In 2016, the DUFB program was expanded to 12 additional supermarkets, for a total of 17 intervention stores. Of those 12 supermarkets, six supermarkets were assigned to the DUFB program based on proximity to pilot stores, relatively higher Electronic Benefit Transfer (EBT) sales, and store demographics. The remaining six supermarkets either self-selected into the program (n=4) or the grocery retailer selected the supermarkets to participate using other criteria (n=2). We used transaction data from 2015 and 2016 in our analyses.

The partner in grocery retailer provided data at the item level grouped by transaction for all transactions in the intervention and control stores for each day in 2015 and 2016. Each transaction includes item; quantity purchased; price, date, time, and location of purchase; method of payment (including EBT, EBT + DUFB, or other); unique loyalty card identifiers; and product description. The loyalty program provided customers with coupons and perks.
This was a quasi-experimental design with an intervention group and a control group that included a 7-month non-intervention period, followed by a 5-month intervention period in both 2015 and 2016.

In 2015, the DUFB program provided a 100% subsidy on all fresh produce to SNAP recipients in the intervention group, which could be redeemed on locally-grown produce up to $20 per day. In 2016, the subsidy was earned on locally-grown produce and the subsidy could be redeemed on all fresh produce up to $20 per day. In both years, the program was only active between August and December. DUFB points could be earned and redeemed until the end of the program; no DUFB points were earned on qualifying produce during the 7-month non-intervention period.

In both years, DUFB points were earned and redeemed automatically when both an EBT card and a loyalty card were used in a transaction. Therefore, we only included loyalty card transactions in our analyses, for a total of 209,775 and 566,649 transactions in intervention stores in 2015 and in 2016, respectively. SNAP recipients earned a point for every whole dollar spent on qualifying produce, and the points could be used 24 hours after the transaction. Customers were made aware of the program via in-store signage, circular letters, and direct mail sent to SNAP participants living in zip codes with intervention stores. DUFB points were highly visible on the receipt after completing a transaction, and cashiers were trained to describe the program to customers.

To control for secular trends using a comparison group, we obtained a list of the remaining stores that did not participate in the DUFB program in 2016 (n=45), and the grocery
retailer approved the release of transaction data from 15 stores of our choice. To select these 15 stores, we used Coarsened Exact Matching (CEM) and a linear probability model. The CEM and linear probability models included five variables: the percentage of total store sales attributed to EBT, the number of employees per store, population density, median income for households receiving SNAP benefits, and percentage African-American population. Sociodemographic characteristics were derived from 2010-2014 5-year estimates from the American Community Survey at the zip-code level.

To account for unknown treatment assignment, we first used CEM to match potential control stores to intervention stores that self-selected (n=4) or were selected using other criteria (n=2), which produced three viable matches (Iacus et al., 2011b). Then we used a linear probability model to create a scoring function for all stores located within 140 kilometers of the two pilot stores, based on visual inspection of the nearest group of stores; and the top 12 stores by score value were allotted to the comparison group, for a total of 15 control stores in 2016 (n=761,408 loyalty card transactions). In 2015, the comparison group also included data from the 12 stores that later adopted the program in 2016, for a total of 27 control stores (n=1,073,211 loyalty card transactions). The trends in the percentage of transactions made with a loyalty card per month in intervention and control stores were approximately equal (Supplementary Figure 2.5), suggesting that the quality of matching was high.

Our primary outcome was the percent of total dollars spent on all fresh produce per transaction, which we calculated by summing dollars spent on all fresh produce per
transaction and dividing the sum by the total transaction amount. To identify any potential spillover effects on unhealthy food items, we also evaluated the impact of the DUFB program on spending on sugar-sweetened beverages (SSBs). Thus, our secondary outcome was the percent of total dollars spent on SSBs per transaction.

We used a difference-in-difference (DiD) model to estimate the average change in spending on all fresh produce (or SSBs) per week between the pre-intervention period and the intervention period among transactions made by SNAP recipients. Transactions made by a SNAP recipient were defined as transactions where an EBT card was used. Independent variables included experimental condition (Intervention); time period (Time); and an Intervention x Time interaction to capture difference-in-difference changes over time.

\[
y_i = \beta_0 + \beta_1 (\text{Intervention}) + \beta_2 (\text{Time}) + \beta_3 (\text{Intervention} \times \text{Time}) + \epsilon_i
\]

In our primary analysis, we used a difference-in-difference-in-difference (DiDiD) model to estimate the average change in spending on all fresh produce (or SSBs) per week between SNAP and non-SNAP transactions. Thus, we added an Intervention x Time \times SNAP interaction to capture DiDiD effects. The model also includes store fixed effects and quarter fixed effects to control for seasonality.
\[ y_i = \beta_0 + \beta_1 (\text{Intervention}) + \beta_2 (\text{Time}) + \beta_3 (\text{SNAP}) \]

\[ + \beta_4 (\text{Intervention} \times \text{Time}) + \beta_5 (\text{Intervention Store} \times \text{SNAP}) + \beta_6 (\text{Time} \times \text{SNAP}) \]

\[ + \beta_7 (\text{Intervention Store} \times \text{Time} \times \text{SNAP}) + \beta_8 (\text{Store Fixed Effect}) \]

\[ + \beta_9 (\text{Quarter Fixed Effect}) + \epsilon_i \]

To control for possible endogeneity between household characteristics and DUFB program participation, we also ran models with loyalty card fixed effects. These analyses were conditional on having at least one loyalty card transaction in both the pre-intervention and intervention periods (n=764,644 and 821,560 transactions in 2015 and 2016, respectively).

To account for differences in produce eligibility, all analyses were stratified by year. All analyses were conducted using R.

2.3 Results

In 2015, the average percentage of transactions made with an EBT card and a loyalty card per month was higher in intervention stores (14.2%) compared to control stores (7.6%) during the DUFB program; and the average percentage of transactions made with an EBT card and a loyalty card per month during the DUFB program was 9.5% in intervention stores and 6.3% in control stores in 2016. In both years, the average percentage of transactions made with an EBT card and a loyalty card in the period before the DUFB program was similar to the percentage during the DUFB program.
Table 2.1: Average Percentage of Total Dollars Spent on all Fresh Produce and Sugar-Sweetened Beverages Per Month

<table>
<thead>
<tr>
<th></th>
<th>2015 Before DUFB</th>
<th>2015 During DUFB</th>
<th>2016 Before DUFB</th>
<th>2016 During DUFB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fresh Produce</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAP</td>
<td>8.41 (0.6)</td>
<td>7.86 (0.36)</td>
<td>8.37 (0.53)</td>
<td>7.41 (0.43)</td>
</tr>
<tr>
<td>Non-SNAP</td>
<td>10.46 (0.52)</td>
<td>9.12 (0.43)</td>
<td>11.1 (0.63)</td>
<td>9.75 (0.34)</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAP</td>
<td>8.65 (0.59)</td>
<td>7.56 (0.34)</td>
<td>8.94 (0.63)</td>
<td>7.76 (0.35)</td>
</tr>
<tr>
<td>Non-SNAP</td>
<td>11.38 (0.61)</td>
<td>9.92 (0.3)</td>
<td>11.42 (0.61)</td>
<td>9.91 (0.3)</td>
</tr>
<tr>
<td><strong>SSBs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAP</td>
<td>4.24 (0.18)</td>
<td>4.09 (0.18)</td>
<td>4.45 (0.3)</td>
<td>4.34 (0.19)</td>
</tr>
<tr>
<td>Non-SNAP</td>
<td>2.72 (0.16)</td>
<td>2.66 (0.14)</td>
<td>2.74 (0.23)</td>
<td>2.71 (0.13)</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAP</td>
<td>4.54 (0.28)</td>
<td>4.35 (0.21)</td>
<td>4.75 (0.32)</td>
<td>4.62 (0.25)</td>
</tr>
<tr>
<td>Non-SNAP</td>
<td>2.71 (0.2)</td>
<td>2.62 (0.14)</td>
<td>2.78 (0.21)</td>
<td>2.77 (0.16)</td>
</tr>
</tbody>
</table>

Format: "mean (sd)"

2.3.1 Fresh Produce

In 2015, the average percentage of total dollars spent on all fresh produce was lower during the DUFB program period (8.1%) than the period before the DUFB program (8.4%) among SNAP recipients shopping in intervention stores (Table 2.1). Similarly, spending on all fresh produce was lower during the DUFB program period (7.7%) compared to the period before the DUFB program (8.6%) among SNAP recipients shopping in control stores (Figure 2.1). However, the DiD model showed that the decrease in spending on all fresh produce during the DUFB program period was 0.6 percentage points smaller in intervention stores (-0.4%) than control stores (-1.0%) (p<0.001) (Table 2.2), which is equivalent to a 5.0% increase in spending on fresh produce in the intervention group.

In 2016, the average spending on all fresh produce was also lower during the DUFB program period (7.5%) compared to before (8.2%) among SNAP recipients shopping in intervention stores (Figure 2.2). In the DiD model, the difference in spending on all fresh produce before and during the DUFB program period was 0.2 percentage points
smaller in intervention stores (-0.80%) versus control stores (-0.09%) among SNAP recipients (p<0.001) (Table 2.3), which is a 1.6% increase in spending on fresh produce in the intervention group.
Table 2.2: Model Estimates for the Effect of the Double Up Food Bucks (DUFB) Program on the Average Percentage of Total Dollars Spent, 2015

<table>
<thead>
<tr>
<th>Item</th>
<th>Intervention stores (n=5)</th>
<th>Control stores (n=27)</th>
<th>DiD</th>
<th>DiDiD</th>
<th>DiDiD (†)</th>
<th>DiDiD (‡)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>During</td>
<td>Diff.</td>
<td>Before</td>
<td>During</td>
<td>Diff.</td>
</tr>
<tr>
<td>Fresh Produce</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAP</td>
<td>8.40%</td>
<td>8.10%</td>
<td>-0.40%</td>
<td>8.60%</td>
<td>7.70%</td>
<td>-0.90%</td>
</tr>
<tr>
<td>non-SNAP</td>
<td>10.40%</td>
<td>9.30%</td>
<td>-1.10%</td>
<td>11.00%</td>
<td>9.80%</td>
<td>-1.20%</td>
</tr>
<tr>
<td>SSBs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAP</td>
<td>4.80%</td>
<td>4.70%</td>
<td>-0.10%</td>
<td>5.00%</td>
<td>4.80%</td>
<td>-0.20%</td>
</tr>
<tr>
<td>non-SNAP</td>
<td>3.10%</td>
<td>3.00%</td>
<td>-0.10%</td>
<td>3.00%</td>
<td>2.80%</td>
<td>-0.10%</td>
</tr>
</tbody>
</table>

DiD = difference-in-difference
DiDiD = difference-in-difference-in-difference
(†) Model estimated with store fixed effects and quarter fixed effects
(‡) Model estimated with loyalty card fixed effects
*p<0.05; **p<0.001
Table 2.3: Model Estimates for the Effect of the Double Up Food Bucks (DUFB) Program on the Average Percentage of Total Dollars Spent, 2016

<table>
<thead>
<tr>
<th>Category</th>
<th>Intervention stores (n=17)</th>
<th>Control stores (n=15)</th>
<th>DiD</th>
<th>DiDiD</th>
<th>DiDiD (†)</th>
<th>DiDiD (‡)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>During</td>
<td>Diff.</td>
<td>Before</td>
<td>During</td>
<td>Diff.</td>
</tr>
<tr>
<td>Fresh Produce</td>
<td>8.20%</td>
<td>7.50%</td>
<td>-0.80%</td>
<td>8.70%</td>
<td>7.80%</td>
<td>-0.90%</td>
</tr>
<tr>
<td>SNAP</td>
<td>10.40%</td>
<td>9.30%</td>
<td>-1.10%</td>
<td>11.10%</td>
<td>9.80%</td>
<td>-1.30%</td>
</tr>
<tr>
<td>non-SNAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSBs</td>
<td>4.90%</td>
<td>4.80%</td>
<td>-0.10%</td>
<td>5.10%</td>
<td>5.00%</td>
<td>-0.02%</td>
</tr>
<tr>
<td>SNAP</td>
<td>3.10%</td>
<td>3.10%</td>
<td>-0.02%</td>
<td>3.00%</td>
<td>2.90%</td>
<td>-0.04%</td>
</tr>
<tr>
<td>non-SNAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DiD = difference-in-difference
DiDiD = difference-in-difference-in-difference
(†) Model estimated with store fixed effects and quarter fixed effects
(‡) Model estimated with loyalty card fixed effects
*p<0.05; **p<0.001
In 2015, the DiDiD model controlling for store and quarter fixed effects showed that the decrease in spending on all fresh produce was 0.5 percentage points smaller among SNAP versus non-SNAP recipients (p<0.001) (Table 2.3). Results were similar in a model with loyalty card fixed effects (DiDiD=0.4%; p<0.001). In the 2016, the decrease in spending on all fresh produce was 0.2 percentage points smaller among SNAP versus non-SNAP recipients (p<0.001). We observed similar results from a model with loyalty card fixed effects (DiDiD=0.1%; p=0.01).

Figure 2.1: Percentage off Total Dollars Spent on all Fresh Produce Per Month Among SNAP Recipients Shopping in Intervention and Control Stores, 2015
2.3.2 Sugar-Sweetened Beverages

In intervention stores in 2015, the average percentage of total dollars spent on SSBs was not meaningfully different during the DUFB program period (4.7%) than the period before the DUFB program (4.7%) among SNAP recipients shopping in intervention stores (Table 2.1). The difference was similarly negligible among transactions in control stores. In the DiD model, the difference in spending on SSBs between baseline and the DUFB program period did not meaningfully differ in intervention stores compared to control stores (Figure 2.3). We observed similar estimates in 2016 (Table 2.3).

The DiDiD model controlling for store and fixed effects showed no difference in
the effect of the DUFB program on SSB spending between SNAP and non-SNAP recipients (p=0.06). Results from a model with loyalty card fixed effects were similar (p=0.98) (Table 2.2). We also observed no significant DiDiD effects in 2016 (Figure 2.4).

![Figure 2.3: Percentage of Total Dollars Spent on all Fresh Produce Per Month Among SNAP Recipients Shopping in Intervention and Control Stores, 2016](image)

2.4 Discussion

Though the percentage of spending on fresh produce was lower during (versus before) the DUFB program period in all stores, we observed a meaningfully smaller decrease in spending in intervention stores compared to control stores, indicating that the program had a positive effect on sales of fresh produce during the program period. The results suggest that the percentage increase in produce sales among the SNAP recipients was about 4% and 1% higher than baseline in 2015 and 2016, respectively. Though the impact was small,
our findings are promising because increased intake of fruits and vegetables, in conjunction with decreased consumption of energy-dense foods, has been shown to prevent obesity and weight gain (Mytton et al., 2014).

In a recent study, researchers tested a double-dollar pricing incentive on eligible fruits and vegetables in a supermarket in a low-income rural Maine community and found that weekly spending on fresh fruits and vegetables increased by 53% among SNAP participants who redeemed coupons (Polacsek et al., 2018). The effect size of this intervention was large, perhaps because the investigators assisted participants with joining the store’s loyalty card program and sent monthly messages to remind participants to use their loyalty card. In the evaluation study of the Healthy Incentives Pilot (HIP), a price rebate of 30-cents for ev-
ery SNAP dollar spent on eligible fruits and vegetables was shown to increase spending on eligible fruits and vegetables by 11% in SNAP households (Bartlett et al., 2014). The larger impact on spending may have been due to recruitment of SNAP households (versus stores) and training HIP participants on procedures. Though previous studies have shown similar effects of produce incentives on fruit and vegetables purchases, our evaluation suggests that the impact of the DUFB program was small, potentially due to the focus on locally-grown produce or imperfect awareness and understanding of the DUFB program (Young et al., 2013; Bowling et al., 2016; Ratigan et al., 2017; Polacsek et al., 2018; Bartlett et al., 2014).

To increase awareness, future programs should invest in participant outreach and education. Although we observed significant increases in fruit and vegetable spending, our findings suggest that the DUFB program had no significant effect on SSB purchases. Though spending on SSBs did not meaningfully change, it is possible that the shift in spending on produce was the result of changes in spending on other unhealthy or healthy food items, including produce not eligible for the incentive (e.g., frozen fruits and vegetables) and non-grocery items. Though we cannot identify where the shift occurred from our results, exploratory analyses from the HIP study suggest that increases in fruit and vegetable intake substituted intake of refined grains (Olsho et al., 2016). A recent review of food pricing interventions reported that multicomponent interventions appear to be the most effective in improving multiple diet behaviors (Afshin et al., 2017). For example, researchers found that SSB purchases decreased by 80-cents per week in response to a 30% financial incentive in conjunction with a restriction on the purchase of SSBs (French
These results suggest that the effects of the DUFB program on improving the whole diet may be more effective with a combination of price incentives and restrictions.

Our study had many strengths, including our large sample size with millions of individual transactions over two years. In particular, our focus on multiple supermarkets in low-income areas was novel, and the expansion of the DUFB program to supermarkets is expected to significantly reach more SNAP participants. The 100% subsidy was also larger than incentives used in several previous studies. However, our study also had several limitations. The definition of “fresh” produce was not always clearly defined in supermarkets, which may have been a source of confusion for consumers. We were also not able to estimate the impact of the program on changes in consumption or absolute changes in fresh produce purchases because sales data were proprietary. Finally, the grocery retailer intentionally selected supermarkets with the highest EBT sales to participate in the DUFB program, which were located in relatively homogenous geographic locations. This approach reduced the burden of DUFB program implementation, but it also artificially inflated population density and the percentage African American in the intervention group.

In this evaluation study, we observed a small, positive effect of a DUFB program on fresh produce purchases among low-income consumers shopping in supermarkets. The results highlight the effectiveness and feasibility of dollar matching programs, and provide support to the USDA to expand this provision. Future studies should analyze the effect of the DUFB program in 2017 and 2018, when the subsidy was both earned and redeemed on all fresh produce. Even though effects on frozen and canned vegetables were smaller in a
similar study, expanding the pool of eligible produce to include frozen and canned produce may be a viable option for continuing the program in other seasons. Efforts to increase the awareness and understanding of the program would also be valuable (Polacsek et al., 2018).

Figure 2.5: Percentage of Transactions Made with a Loyalty Card per Month in Intervention and Control Stores, 2016
Chapter 3

Testing Behavioral Economic Nudges in a Chain of Convenience Store to Determine Impact on Healthy Food Choice Behavior

3.1 Introduction

The fields of economics and psychology have been closely linked since the late eighteenth century (Ashraf et al., 2005). One of the standard economic theories used to explain human behavior and decision making is “Rational Agent Theory” (Rachlin, 2003). According to Rational Agent Theory, humans are fully rational beings, and a human’s decision to perform a behavior (or not) is based on a simple mental calculation of risks versus benefits (Roberto and Kawachi, 2014). If the benefits outweigh the risks, she or he will perform the behavior (Roberto and Kawachi, 2014). In this theory, humans are seen as utility-maximizers, and will behave in ways that bring them the most benefit and utility (Rachlin, 2003). We know this is not always the case, however (Rachlin, 2003). Traditional economics and Rational Agent Theory do not explain everything that goes into human behavior and decision making (Rachlin, 2003). This is because people are not fully rational, but boundedly rational (Rachlin, 2003). Instead of mentally calculating the risks versus benefits for all of our decisions, humans take “reasoning shortcuts” to make decisions (Rachlin, 2003). These “reasoning shortcuts” often lead to suboptimal decision-making (Roberto and Kawachi, 2014). For
example, a “reasoning shortcut” is placing present needs before future needs. Behavioral economics provides a road map to recognizing these “reasoning shortcuts” and fixing human decision errors (Roberto and Kawachi, 2014).

Convenience stores are a key location for the provision of food in communities characterized by a lack of full service grocery stores. These stores traditionally carry predominantly indulgent food products and fewer healthier options. These retail settings, however, can be used to nudge healthier food choice through the above “reasoning shortcuts”. Research has been conducted suggesting that using behavioral economic strategies can encourage healthier food choices in grocery stores (Foster et al., 2014). Similar behavioral economic nudges could prove successful in convenience store settings, which is what we set out to test.

Prior research has been conducted exploring how behavioral economic strategies can be used in the retail food environment to encourage healthy food choice. For example, we know that a number of marketing, product placement, and other factors can influence buyers’ decision making (Roberto and Kawachi, 2014; Riis, 2014). Marketplace “cues”, such as appealing packaging or attractive displays at checkout, may increase the likelihood that consumers will purchase a product (Richards and Sindelar, 2013). Product placement strategies such as placement on endcaps to feature promoted items and thus encourage their purchase, have been associated with increased sales (Hui et al., 2013). Using signage to denote healthier food options has also been shown to increase sales in convenience stores (Dannefer et al., 2012).
We set out to test three behavioral economic nudges in a set of 32 convenience stores in the Research Triangle region of North Carolina. Our aim is to evaluate the effectiveness of these behavioral economic nudges on store sales of targeted items.

3.2 Methods

3.2.1 Study Design

We partnered with a chain of convenience stores located in one state in the American South to conduct a field study in 32 of those stores located in a one mid-sized city and one adjacent college town. Stores were split into 16 pairs of matched stores based on sales, location, and customer demographics in collaboration with the President of the franchise. Each pair was then randomized into Group A or Group B, which would receive a different set of experiments, using a random number generator. Then, a list generator was used to randomize the order of the pairs and a random number generator was used to randomized each pair so that one store was a treatment and one was a control. A 2x2 methodology was employed, with all stores organized into two treatment and two control groups with eight stores in each arm. IRB approval was required because research for, and the implementation of, two of the experiments involved human subjects. Amazon’s Mechanical Turk was used to select questions for two of the experiments and one experiment induced in-person interactions with human subjects. Approval for the study was provided by the Duke University Institutional Review Board, and the UNC IRB agreed to rely on the Duke Institutional Review Board.
3.2.2 Intervention

In planning the intervention, the President of the franchise was presented with a description of fifteen different behavioral economic strategies to test in his stores. After review and discussion, the study team, in collaboration with the President, ultimately selected three experiments to implement. We field study tested these three experiments (Ask, Ask + Pump, and Bundling) in those stores. The three experiments are described below. Prior to the implementation, each intervention was tested in one selected store for a four-week run-in period. This pilot period allowed each intervention to be tested and changes made before full implementation.

3.2.3 Ask Experiment

By changing the context in which individuals make a decision, behavioral scientists have shown that outcomes of a decision can be altered (DellaVigna, 2009; Thaler and Sunstein, 2009; Thaler, 1980; Kahneman et al., 1991; Johnson et al., 2012). Both the Ask and Pump (description follows) experiments are based on the idea of encouraging individuals to make healthier decisions. The Ask experiment was implemented in eight of the stores, which also had a matched control. In this experiment, store clerks were instructed to ask all customers making purchases at the cash register “Would you like to buy a banana today?”. This behavioral nudge is known as “prompted choice” (Sunstein, 2014). It presents the customer with a positive, non-intrusive choice to accept or decline that, absent the experiment, they would otherwise not have encountered.
This message was chosen based on the testing of several options using Amazon Mechanical Turk (mTurk). The Amazon mTurk platform is one of the first and largest platforms for crowdsourcing work, which is a method of obtaining information from large groups of people—particularly online (Sheehan and Pittman, 2016). Amazon’s mTurk is a platform where “requesters” (the researchers in this case) post online human intelligence tasks (HITs) for “workers” to complete (Sheehan and Pittman, 2016). Within mTurk, participants, who are paid a nominal rate ($0.35), were asked to rate the various message options on whether each message would be likely to persuade them to buy a particular item. The average completion time was 5 minutes. For our survey, participants must have had at least 97% of prior HITs on mTurk approved.

All stores, treatment and control, received training sessions administered by training staff at the chain’s headquarters. These training sessions were designed to increase banana sales. Role playing was used to practice the Ask at the point of sale, using the headquarters’ mock checkout counter and display. All in-store staff, from cashiers to managers to store owners, were asked by management staff to “pledge” the amount of bananas they thought they could sell weekly during this Ask experiment. Staff were then asked to sign the pledge as a means of committing to selling the amount of bananas they listed on the form. Reminder cards with the selected Ask question were distributed to all treatment stores and displayed behind the cash registers so that clerks may view them. The duration of this experiment was 2 months.
3.2.4 Pump Experiment

In this experiment, the research team wanted to evaluate whether a message displayed to customers who purchased gas would influence banana sales in the stores. The following message was displayed once customers had selected their desired fuel grade and began to pump their gas:

“For a burst of energy, buy a banana inside. Loaded with nutrients and energy.”

This nudge is a positive informational message. The message shares healthy and positive information about bananas and also where to buy them. To select the above messaging, an online survey was again administered to a lay audience using the platform mTurk. Participants were provided a pairwise comparison of various messages. This message was non-animated, displayed on the majority of the screens at the treatment stores throughout the month and was cycled along with other promotional messages during the duration of the time needed to pump gas.

The Pump experiment occurred in tandem with the Ask experiment for one study arm in eight stores. Prior to implementing the Pump experiment, the Ask experiment was run independently for two weeks.

3.2.5 Bundling Experiment

Behavioral scientists have employed different marketing techniques to influence consumers’ purchases such as product placement and bundling, which plays a vital role in the food retail environment, as such techniques increase the convenience and salience of the targeted product (Glanz et al., 2012; Chance et al., 2014; Sunstein, 2014). Thus, the bundling experiment
aimed to test these strategies with healthy products in our convenience stores. Specifically, we chose to test the effect of displaying two healthier items together that have a popular understanding of being consumed at the same time to appeal to individuals seeking variety (Harris, 2006). A team of nutrition staff at UNC were provided with a list of all items available in our stores and two “bundles” of items were identified as healthy and popular: yogurt and granola and bananas and peanut butter. This experiment was conducted in 16 treatment stores with matched controls and these stores had the items placed together physically. In addition, the research team developed signage that was placed next to these items in one treatment arm (eight stores). This signage contained the messages “Better together: Try yogurt and granola today” or “Better together: Try bananas and peanut butter today” on each respective display. These messages were also determined using an mTurk online survey with a lay audience using a pairwise comparisons of various messages. This experiment had a duration of 2 months.

All experiments were monitored for implementation fidelity through a series of observations conducted by research staff. These checks were executed at a randomized schedule for the “Ask” and “Pump” experiments, with a variety of days of the week and times of day represented because staff members change and randomly conducting observations assured us that different store staff would be observed. The team also checked to ensure that none of the control stores were asking customers to buy bananas. The fidelity checks for the “Bundling” experiment were not randomized, but also served as an opportunity for research staff to repair any lost or damaged signage alongside ensuring fidelity to
the bundling protocols.

All communications to store staff were disseminated through the President and the channels of communications used to regularly send information to the individual stores. All stores during the period of these experiments were displaying bananas next to each stores’ cash registers. Bananas were replenished once a week at each store by a contracted fruit company vendor.

3.3 Analysis Plan

3.3.1 Outcome measure

The measurement component to this study was total daily banana sales. Bananas are sold either as singles banana or in 2-packs. An occurrence of either sale is counted as 1 banana sale. That is, if three single bananas were purchased and four banana 2-packs, this would be counted as six banana sales despite 11 actual bananas being sold.

The motivation behind this definition was to frame the customer’s decision to purchase a banana as a discrete choice. The interest is not the quantity of bananas a customer decides to buy, but whether the customer decide to purchase any bananas at all. Put another way, the nudges are considered equally “successful” should they move a customer to buy a single banana or a 2-pack.

The franchise provided four months of data (1st of August through end of November) from 2015 and 2016 (eight months of data, four from each year). Prices for the single and 2-pack remained fixed across both years at $0.89 and $1.59, respectively.
An important assumption is required given the nature of the data. The model assumes that within a given day each single and 2-pack purchased corresponds to a different person; the same person does not purchase bananas twice on two separate visits in one day. This assumption is important because only aggregated sales data were provided. It is impossible to determine if, for example, three single bananas were purchased by the same customer during one visit to the register. This behavior is certainly possible but the assumption is that it occurs very infrequently. The model also assumes that customers are sufficiently rational and price sensitive to, more often than not, elect buying a banana 2-pack versus two single bananas (saving $0.19).

3.3.2 Baseline Data

Figure 3.1 displays weekly banana sale counts for the Ask / Ask + Pump stores. This figure displays the variation in banana sales between stores and across time. In 2015, banana sales were more stable across all stores and more similar in magnitude. In 2016, banana sales were more volatile and, on average, higher. Individual stores also varied in banana sales between consecutive years. In short, Figure 3.1 displays the need for proper controls, like year-month and store fixed effects. Without such controls, average banana sale counts would be biased upwards, inflating the impact—if any—of the Ask and Ask + Pump treatments.
Two estimation methods were used, difference-in-differences (DD) and ANCOVA with baseline values. It is standard in economic analyses to use difference-in-differences, but estimates come at a cost of power (McKenzie, 2012). Given random assignment and a dependent variable (i.e. banana counts) with low auto-correlation, ANCOVA models produce similar estimates to difference-in-differences but without sacrificing as much statistical power.

Difference-in-difference estimation models with fixed effects are suitable given the nature of the experimental design. Store groups move in and out of being either a “treated” store or a “control” store during 2016. The exception are the eight stores in the perpetual control group. This phasing of stores in and out of “treated” status facilitates...
comparison between groups over-time during 2016; the 2015 data provides a baseline pre-treatment period for all groups. Differences-in-difference models provide a framework to estimate treatment effects in a single equation as follows:

\[ y_{it} = \alpha_i + DOW_t + YM_t + \beta_1 SALES_{it} + \beta_2 GAS_{it} + \sum_{j=1}^{4} \theta_j EXPER_{ijt} \]

\[ + \delta_1 A_{it} + \delta_2 P_{it} + \delta_3 B_{it} + c_i + u_{it} \]

The following control covariates are included in the DD model: store-level fixed effects, day-of-the-week and year-month dummies, and daily total grocery and gas sales. Store-level fixed effects control for any unobservable store-level characteristics that do not vary over time. For example, a store’s geographic location may influence customer arrival rates. The average impact a store’s location may have on its banana sales is captured by fixed effects.

Day-of-the-week controls and year-month controls to capture within-week cyclically and seasonal effects were also used. A daily time fixed effect was considered but there are insufficient degrees of freedom in the panel models to allow their inclusion. Capturing the within-week cycles with day-of-the-week variables and month-to-month seasonality produced very similar estimates in the cross-sectional models but required giving up substantially fewer degrees of freedom. They also kept the model from over-fitting the data.

Total daily grocery sales (SALES in dollars) and gas sales (GAS in gallons) were included to control for changes in banana counts due to general increases in economic activity. Increased economic activity could impact the rate at which customers were exposed
to the experiments. For example, cheap gasoline prices may bring in more customers to get gas, increasing the frequency that customers are exposed to the gas pump message nudge.

The remaining dummy variables comprise the core of the difference-in-differences framework. These dummy variables indicate which days each experiment is active and store treatment status (EXPER). The dummy variables denoting when a store is in the Ask (A), Ask + Pump (P), or Bundling (B) experiment produce the coefficient estimates of interest.

The ANCOVA model is provided as a comparison point. Given the experimental design, the ANCOVA model should produce similar results as the DD model. The advantage of the ANCOVA model is that it is more efficient—specifically, that it requires fewer controls—than the DD model (McKenzie, 2012). It also provides more power when dependent variable auto-correlation (banana sale counts) is low, like in our sample. The ANCOVA equation follows:

\[ y_{it} = \bar{Y}_{i,PRE} + DOW_t + YM_t + \beta_1 \text{SALES}_{it} + \beta_2 \text{GAS}_{it} \]

\[ + \delta_1 A_{it} + \delta_2 P_{it} + \delta_3 B_{it} + c_i + u_{it} \]

The covariates for the ANCOVA model are almost identical to the DD model except the baseline covariate (“Y bar”) and a store-level random effect is exchanged for the store-level fixed effect. The baseline covariate is the average daily banana sale counts per store calculated across all days from the pre-treatment period (August - November 2015 and the first 3 weeks of 2016).

Of likely concern is the data generating process for banana sale counts. Banana
Table 3.1: Summary of Banana Sale Counts

<table>
<thead>
<tr>
<th>Store–Days</th>
<th>SD</th>
<th>Zeros</th>
<th>Mode</th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5612</td>
<td>4.24</td>
<td>494</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>5.07</td>
<td>7</td>
<td>29</td>
</tr>
</tbody>
</table>

*Note:*
(Excludes Sales from Treated Stores During ‘ASK’ or ‘ASK\_PUMP’ 2016 Experiments)

sale counts are discrete and non-negative i.e. zero or positive. However, ordinary least squares (OLS) will produce consistent estimates regardless of the data generating process (Matloff, 2017). Standard error estimates can suffer, possibly affecting hypothesis testing, but this is generally only a concern in small sample estimates.

### 3.4 Results

There is statistically significant evidence that the Ask experiment increased daily banana sale counts by at least 1 sale more per day across all treatment stores. The effect size remained roughly the same—1 banana sale more per day—during the Ask + Pump experiment. This is evidence that the inclusion of the informational message at the Pump did not impact banana sales. There is no statistically significant evidence that the Bundling experiment increased banana sale counts. There is, in fact, weak (statistically insignificant) evidence that the Bundling experiment decreased banana sale counts (data not shown). See Tables 3.2 and 3.3 for regression results.

Daily banana sale counts are generally low. Table 3.1 contains the summary statistics for all 2015 and 2016 daily banana sale counts excluding any counts from stores where the Ask and Ask + Pump experiments were run. The mean of this sample of daily banana sale count data is 5.07, implying a mean shift equal to 1 would result in an average of 6
banana sale counts per day (20%).

In other words, were this convenience store chain to implement the Ask intervention (prompted choice for a banana purchase) as store policy across all stores, the long run average of daily banana sale counts could jump from 5 to 6. Note that this assumes the effects of the Ask experiment does not diminish over time. Many behavioral nudges have strong short-term effects that do not persist long-term. However, it is not possible to test the persistence of either experiment given our design.

### 3.4.1 Fidelity Checks

For the Ask experiment, no clerks were observed asking customers to buy bananas during the randomized checks.\(^1\) The pumps, however, always displayed the messages during this portion of the experiment. During the bundling experiment, our team found that the items were always stocked, items were visible, and the signage was correct and present.

### 3.4.2 Change in Sales

Table 3.2 contains the coefficient estimates of the difference-in-differences (DD) models. Results for models without fixed effects are included for comparison. The Ask experiment is measured to have increased banana sale counts by about 1.52 counts—or at least 1 banana count given transactions occur only in discrete values. The same DD model was run with total banana sales in dollars as the dependent variable.\(^2\) This corresponds to an estimated

---

\(^1\)Attempts were made by my coauthors contact the research assistant to get a log of all the visits but were unable to make contact.

\(^2\)Tables excluded.
Table 3.2: Difference-in-Differences, Average Marginal Effects (2015 and 2016)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>OLS</th>
<th>OLS FE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b/se</td>
<td>b/se</td>
</tr>
<tr>
<td>ASK</td>
<td>1.127***</td>
<td>1.516**</td>
</tr>
<tr>
<td></td>
<td>(0.325)</td>
<td>(0.596)</td>
</tr>
<tr>
<td>ASK_PUMP</td>
<td>1.270***</td>
<td>1.185**</td>
</tr>
<tr>
<td></td>
<td>(0.301)</td>
<td>(0.500)</td>
</tr>
<tr>
<td>BUNDLING</td>
<td>1.193***</td>
<td>-0.244</td>
</tr>
<tr>
<td></td>
<td>(0.265)</td>
<td>(0.332)</td>
</tr>
</tbody>
</table>

significance: *, ** = 0.05, *** = 0.01

Table 3.3: ANCOVA, Average Marginal Effects (2016)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>OLS</th>
<th>OLS RE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b/se</td>
<td>b/se</td>
</tr>
<tr>
<td>ASK</td>
<td>1.245***</td>
<td>1.485**</td>
</tr>
<tr>
<td></td>
<td>(0.283)</td>
<td>(0.656)</td>
</tr>
<tr>
<td>ASK_PUMP</td>
<td>0.977***</td>
<td>1.184***</td>
</tr>
<tr>
<td></td>
<td>(0.256)</td>
<td>(0.284)</td>
</tr>
<tr>
<td>BUNDLING</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>(0.203)</td>
<td>(0.313)</td>
</tr>
</tbody>
</table>

significance: *, ** = 0.05, *** = 0.01

$1.74 increase in banana sales. Both estimates are statistically significant at the 5% level.

The DD estimates for banana sales in counts and dollars due to the Ask + Pump correspond to 1.19 and $1.37, respectively. Like the Ask experiment, the Ask + Pump estimates are statistically significant at the 5% level. The effect sizes are similar to those of the Ask experiment. It is therefore reasonable to conclude that the Pump component did not contribute to increasing sales.

Unlike Ask or Ask + Pump, the DD estimates for the Bundling experiment were not statistically significant. The estimates were -0.25 and -$0.30. Assuming they had been significant, their negative value would have been curious and counter-intuitive. It is unlikely
that the bundling of bananas, given the option to buy bananas unbundled, would lead to
dlower sales. What is more likely is that the bundling experiment had no effect on sales.

Table 3.3 contains coefficient estimates of the ANCOVA models. The ANCOVA
results were very similar to the DD results. For the Ask, Ask + Pump, and Bundling exper-
iments the counts and dollars were 1.49 and $1.71, 1.18 and $1.35, and 0.025 and $0.01,
respectively. Like the DD results, the Ask and Ask + Pump estimates are statistically sig-
nificant and the Bundling results are not.

3.5 Discussion

We set out to test the effectiveness of three behavioral economic nudges in a set of 32 con-
venience stores on sales of healthier foods. We found the Ask (prompted choice) nudge re-
sulted in the intended behavior, an increase in sales of the promoted item, while the Bundling
and Pump experiments did not increase sales of the promoted items. Specifically, the Ask
experiment resulted in an increase in the promoted item, bananas, by roughly one more
additional banana sale per day. To put this in the context of typical sales volume for this
product in this type of store, the increase in sales that we saw equates to a 20% increase in
average counts across the store chain.

The results are not trivial. Behavioral economics nudges should be seen as a way
to solve the last mile problem (i.e. how to change behavior at the margin) and not to trans-
form individuals into something they are not. Behavioral economics strategies are about
making customers take an extra step that they would not have taken otherwise. As such,
given the low sales base for bananas in these stores in general, the results are promising. There is no reason to believe that this nudge would not work proportionally with a much higher sales volume class of products. These results, combined with those of other studies (Baca-Motes et al., 2013; Liu et al., 2018; Hui et al., 2013; Holmes et al., 2012) can be used to reassure store managers that prompting customers to buy healthy items can be beneficial from a business perspective.³ It is also possible to conclude that the prompted choice nudge in particular could be generalized to other food retail environments. For example, grocery stores clerks could suggest the purchase of a healthy product at checkout. The financial cost of such an intervention would be low. The prompted choice nudge, however, should be done with considerations for the comfort and sanity of both the staff and customer, each of which can grow tired and annoyed with repeatedly asking or hearing the same question.

3.5.1 Limitations

As with any study, there were limitations. The research staff did not develop or monitor the trainings the store management staff conducted for in-store staff on the Ask experiment, or create the pledge card. These efforts were initiated by the retailer. They were, however, standardized so that all staff in treatment stores received the same training and materials. In addition, the experiment messages on the pumps were inadvertently piloted in three stores rather than our intended one due to the need to test the message on different types of screens and then failure to remove the messages after testing. Our results also assume that the effects of the Ask and Ask + Pump experiments do not diminish over time. We know, however, that

³This unfortunately also implies that prompted choice of a less healthy and hard to resist product, like candy or soda, could result it equal or greater sales than those observed for bananas.
some interventions employing behavioral economic nudges have strong short term effects that do not persist long term (Cohen et al., 2015). It was not possible to test the persistence of either the Ask or Ask + Pump experiments because the data do not exist. Lastly, seasonality may have had a role in the sales of bananas as fruit is less likely to be purchased in cold months. We did, however, address this by using a study design that employed control stores and year-month controls.

3.5.2 Strengths

While there were limitations, there were also strengths to our study. The close partnership with the company president allowed us access to all 32 stores and their staff, which eased implementation and improved fidelity to the experiment protocols. Further, we had a robust number of stores to work with (32) allowing for both treatment and control stores. This is the study’s biggest strength: the implementation of a near-randomized controlled trial experimental design under real-world conditions. While stores were paired with input from our partner, randomization, implementation, and analysis was done independently. This is uncommon in academic and private sector partnerships despite producing better research.

3.6 Recommendations

In talking with store staff following the Ask experiment, we found that there was significant fatigue with repeatedly asking customers if they wanted to buy a banana. Some store staff reported that they curtailed asking repeat customers because it got to be a nuisance for the customers. This may be solved by promoting different healthy items daily, weekly, or
monthly instead of consistently promoting one item. Further, we only tested one setting and
one item in one product category. Future research should test similar behavioral economic
nudges in other settings such as dollar stores and full-service grocery retails and expand
to other product categories. We tested our nudges during a specific time in the shopping
journey (check-out). Future studies could examine the impact of nudges during different
times such as while browsing the store or upon entering.

3.7 Conclusion

One of three behavioral economic nudges tested—prompted choice—showed promise in
promoting the purchase of healthy foods thus demonstrating the possible benefits of using
these nudges in larger, full-service grocery stores. This study contributes to the dearth of
literature on the impact of use of behavioral economic nudges in a grocery retail setting.
Additional studies are needed to determine whether the same results are garnered in full-
service grocery stores and what other behavioral economic nudges may also be effective.
Conclusion

This dissertation analyzes two novel pilot programs that aim to increase the purchase of fruits and vegetables. Historically, policies and programs aimed at changing purchasing behavior of consumers were passive, like guidelines, or regressive, like taxation. In the last 10 years, policy makers have accelerated experimenting with targeted subsidies and behavioral nudges as possible methods to shift purchasing behavior at the margin. The three chapters of this dissertation explore a small corner of this expanding experimental policy universe, rigorously measuring the efficacy of both pilot programs, and offering suggestions for improvement.

During its infancy, any pilot program can feel destined for success, especially one with a compelling story behind it. Proper experimental design and rigorous econometric analysis helps ensure that story is supported by evidence of success, not just good narrative. This is especially true when trying to measure programs that aim to increase something as notoriously unpopular as fruits and vegetables. The Double Up Food Bucks program, analyzed in Chapter 1 and Chapter 2 of this dissertation, had a wonderful story prior to its 2014 piloted expansion into grocery stores. It started in a farmer’s market and expanded exponentially into farmers markets across the country; it made fresh local produce more affordable to low income households; and it moved more money into the pockets of local farmers. But
the popularity of Double Up Food Bucks in farmers markets did not necessarily imply it would be popular in grocery stores. More importantly, popularity of a program is not evidence that the program has succeeded in completing its stated goal. Likewise, behavioral nudges, like those analyzed in Chapter 3 of this dissertation, have received enough hype and attention recently that any new and quirky behavioral nudge can feel like a winner a priori—even if that the nudge aims to increase banana sales in an environment known most for selling beer, soda, and candy bars (gas station convenience stores).

Ultimately, what ties this dissertation together is a desire to keep my own optimism about promising new interventions in check. I believe policy makers should find non-intrusive and creative ways to help people improve their health and lives. There is consensus in the literature that what Americans (and most of the world) consume is unhealthy in the long run. But a necessary step in changing what folks consumes is changing what folks choose to buy. If there is any hope in changing consumption behavior, policy makers must start influencing purchasing behavior. Proper experimental design and econometric analysis ensures that researchers and policy makers avoid falling victim to new programs and interventions with compelling narratives and instead stay focused on studying and funding those that may actually have a chance at improving people’s lives. The two pilot programs researched in this dissertation show promise: the Double Up Food Bucks incentive program induced a measurable, albeit small, increase in produce purchases, and 1 of the 3 behavioral nudges generate modest increases in banana purchases. The findings are useful, even if not groundbreaking, contributions to a limited, but growing, body of literature about healthy
food choice interventions. I leave it to future academics to continue vetting and expanding on the research explored within this dissertation.
Appendix

Double Up Food Bucks Stores: Motivation for Matching

Initially, not all treated stores were going to be matched to a control. This was due to how the 17 original treated stores were selected. The parent company intentionally selected stores with some of the highest EBT (aka SNAP Electronic Benefit Transfer (EBT) Card) sales that were also within relatively similar geographic locations. This reduced the burden of advertising and implementing DUFB for the parent company. The unfortunate downside of this implementation is that it effectively removed any likely matches for treated stores located in the most Urban areas (e.g. Grand Rapids and Battle Creek).

Here is an example to illustrate why it is infeasible to matching all treated stores and instead expand selection algorithmically on observables. If we calculate the percentage of the population by zip code that is African American then split the data into treatment and control groups, we get the following:

```r
#> Difference in Means (Treated - Control) = 7.848600
```

```r
#>
```

```r
#> Population, % Black (Treated, Top 10):
```

```r
#> [1] 26.06 21.95 19.79 18.88 18.69 14.69 11.86 8.53 5.64 5.64
```

129
What these results tell us is how potentially distinct the populations are within the zip codes containing the treated stores. Sorting population percentages in descending order, no good match exists within the control stores for the top 7 treated stores. One variable is the simplest case; matching only gets more difficult as one brings in more variables to match.

Considering the separation between some of the treated stores and all of the control stores, it was prudent to rethink the store selection and matching strategy.

It must be noted that matching is not a necessary step during every design phase. It is, in large part, a way to hedge against the possibility that merely selecting the next top 15 stores by EBT sales could sour the estimates. Matching a smaller set of treatment stores against a larger pool of controls can often produce estimates less sensitive to even the smallest changes in some model specifications (Imbens and Rubin, 2015). However, other models and tools (like regression) are in relatively unperturbed by a lack of design-phase matching, but still benefit from having a larger sample size (Angrist and Pischke, 2008).

Fortunately, the larger sample size arrived after initial matching strategy was designed and implemented. Data for an additional 14 controls stores were retroactively provided by the parent company. These additional 14 stores came primarily from rural areas similar to the 6 stores excluded from the scoring via Linear Probability Model. This bol-
stered the amount of data available for the control group, increasing the precision of model regression estimates.

**CEM Matching Details**

Like most data-dependent endeavors, the most tedious part of matching the stores was obtaining enough variables. Once enough data were obtained, variables were selected on how best they captured data from the following dimensions:

- Demographics (e.g. race)
- Income/wealth
- Population density (e.g. urban vs rural)
- Store attributes
- Store EBT sales

One may assume that more variables makes matching easier. This is only true insofar as it provides one with a large pool of options. It is still necessary to carefully select how many variables one is using because matching becomes more and more difficult with each added variable used. This is especially true with a small sample size.

The matching covariates that were finally selected are:

- `pct_black`: Percentage of population that is black (zip code level)
- `dens_pop`: The population density (people per square mile, zip code level)
- `income_p50_snap_yes`: Median income for people who have received SNAP or similar assistance (zip code level)
Table 3.4: CEM Match Matrix

<table>
<thead>
<tr>
<th></th>
<th>G0</th>
<th>G1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>44</td>
<td>6</td>
</tr>
<tr>
<td>Matched</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Unmatched</td>
<td>30</td>
<td>3</td>
</tr>
</tbody>
</table>

- **store_n_associates**: The number of associates employed in each store.
- **ebt_sales_pct**: Percentage of total stores sales attributed to EBT/SNAP.

**Results of Match**

G0 represents the “control” group and G1 represents the “treated” group. One can observe that 3 “treated” stores were matched to 14 “control” stores. Each of the 3 treated stores was matched to its closest control store by driving distance.

**Covariate Cut-points**

The CEM procedure depends heavily on the “cut-points” selected for each variable. This is akin to setting the cut-off points when turning a continuous variable into a categorical variable. For example, when converting income values from dollars into low–middle– and high–income groups, at least 4 cut-points are required (2 of which are the maximum and minimum). What the other 2 cut-points are will greatly affect the match. This leads to the question, for example, should the cut-points be 25000 and 100000 or perhaps the median and the top 10%?

For the matches produced, the following cut-points were created.
Understanding why is best explained using a visualization. Below are graphs of the variables `pct_black` and `income_p50_snap_yes` with their corresponding cut-points. The aim of each cut-point is to balance the creation of reasonably sized partitions while still marking obvious shifts in the underlying distribution.

For example, in the first plot (`pct_black`), there are clearly points where the slope dramatically increases — and then spikes — in the percentage of African Americans. But in the second plot, the slope is more gradual, so the partitioning is aimed more at getting
relatively balanced groups.
Software

*R* (R Core Team, 2018) was almost exclusively used for data preparation work and descriptive analysis. It was also used for generating this document. All regression models, however, were run using *Stata 13* (StataCorp, 2013).

- Dynamic document generation
  - *bookdown* (Xie, 2018)
  - *rmarkdown* (Allaire et al., 2019)
  - *knitr* (Xie, 2019)
  - *Cairo* (Urbanek and Horner, 2015)

- Tables and graphics
  - *gt* (Iannone et al., 2019)
  - *xtable* (Dahl et al., 2018)
  - *kableExtra* (Zhu, 2019)
  - *ggplot2* (Wickham et al., 2019a)
  - *extrafont* (Chang, 2014)
  - *outreg2* (Wada, 2014)

- Data preparation
  - *data.table* (Dowle and Srinivasan, 2019)
  - *dplyr* (Wickham et al., 2019b)
  - *forcats* (Wickham, 2019a)
  - *magrittr* (Bache and Wickham, 2014)
  - *Matrix* (Bates and Maechler, 2019)
  - *purrr* (Henry and Wickham, 2019)
  - *readr* (Wickham et al., 2018)
  - *stringr* (Wickham, 2019b)
  - *tibble* (Müller and Wickham, 2019)
  - *tidyr* (Wickham and Henry, 2019)
  - *tidyverse* (Wickham, 2017)

- Regression models
  - *lfe* (Gaure, 2019)
  - *reghdfe* (Correia, 2017)
Bibliography


Mancino, L. and Guthrie, J. (2014). SNAP Households Must Balance Multiple Priorities To Achieve a Healthful Diet. *Amber Waves*, pages 1D,2D,3D,4D,5D,6D,7D,8D,9D,10D.


R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.


StataCorp (2013). *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.


Wickham, H. (2019a). *Forcats: Tools for Working with Categorical Variables (Factors)*. R package version 0.4.0.


