



# Three Experiments about Human Behavior and Legal Regulation

## Citation

Svirsky, Daniel. 2019. Three Experiments about Human Behavior and Legal Regulation. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

## Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029491>

## Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

*Three Experiments about Human Behavior and Legal Regulation*

A dissertation presented

by

Daniel Svirsky

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

March 2019

© 2019 Daniel Svirsky

All rights reserved.

## Three Experiments about Human Behavior and Legal Regulation

### **Abstract**

Each chapter of this dissertation presents the results of an experiment.

Chapter 1 tests whether people engage in information avoidance when making privacy decisions. Participants decide whether to share their Facebook profile data with a survey-taker in exchange for money. When people make a *direct* tradeoff between 50 cents and privacy, roughly 64% refuse to share their Facebook data. However, when participants face a *veiled* tradeoff and must “click to reveal” to learn whether privacy is free or costs 50 cents, only 40% remain anonymous, and 58% of participants did not click to reveal to learn which payment option was associated with privacy. The findings show that even people who would otherwise pay for privacy seem able to exploit strategic ignorance and deal away their data for small amounts of money. The findings suggest that privacy regulations aimed at giving people more information about data choices will be difficult to execute.

Chapter 2 measures race discrimination against Airbnb guests. It finds that applications from guests with distinctively African-American names are 16% less likely to be accepted relative to identical guests with distinctively White names. Discrimination occurs among landlords of all sizes, including small landlords sharing the property and larger landlords with multiple properties. It is most pronounced among hosts who have never had an African-American guest,

suggesting only a subset of hosts discriminate. While rental markets have achieved significant reductions in discrimination in recent decades, the results suggest that Airbnb’s current design choices facilitate discrimination and raise the possibility of erasing some of these civil rights gains.

Chapter 3 measures the effect of warning labels on soda purchasing. Governments have proposed text warning labels to decrease consumption of sugary drinks – a contributor to chronic diseases like diabetes. We field-tested the effectiveness of graphic warning labels (vs. text warning labels, calorie labels, and no labels) and assessed consumer sentiment. The findings show that graphic warning labels reduced the share of sugary drinks purchased in a cafeteria, but text and calorie labels did not. We also find that public support for graphic warning labels can be increased by conveying effectiveness information.

## Table of Contents

|   |    |
|---|----|
| 1. Acknowledgments.....   | vi |
| 2. Internet Privacy and Information Avoidance .....                     | 1  |
| 3. Racial Discrimination in the Sharing Economy.....                    | 33 |
| 4. The Effect of Graphic Warning Labels on Sugary Drink Purchasing..... | 67 |
| 5. Appendix.....  | 96 |

## Acknowledgments

Thank you, Charlotte. For everything. Thank you to Mario Svirsky and Fanny Sosenke, who always modeled curiosity, a love of numbers, kindness, and a desire to leave the world better than they found it.

This dissertation benefited from dozens of wonderful, generous people. Christine Exley always had her door open. The idea for the experimental design in the privacy paper was hers. Matthew Rabin offered constant, careful feedback and support. David Laibson helped at both a high level -- it was his idea to focus on privacy as a research topic -- and a specific level -- he was always eager to discuss the minutiae of experimental design. Ben Edelman was a caring, thoughtful mentor. I enjoyed conversations with a host of other great people along the way: Andrei Shleifer, Emily Oster, David Cutler, Josh Schwartzstein, John Beshears, Jerry Green, Oliver Hart, Oren Bar-Gill, Al Roth, and Mike Luca stand out, as well as other students in my program -- especially Oren Danieli, Talia Gillis, David Martin.

And it's worth repeating. Thank you, Charlotte, for everything.

For Charlotte, Lucas, and Leo



## **Chapter 1: Information Avoidance and Internet Privacy**

## 1. Internet Privacy and Information Avoidance

There is a widespread intuition that when making decisions about privacy, people are inconsistent. People share lots of data; people are angry about corporations collecting their data. This intuition that people are being inconsistent is shared widely enough to have a name -- the privacy paradox (Acquisti, 2015). This phrase has been mentioned in roughly 5,000 scholarly articles between 2010 and 2018.

This paper uses an experiment that heightens this paradox and provides evidence for a novel explanation: information avoidance. Even people who are willing to pay nearly an hour's worth of wages for privacy are also willing to give away their data for small money bonuses if given a chance to avoid seeing the privacy consequences of their choices.

In the experiment, participants who complete a survey decide whether to do the survey anonymously or after logging in with their Facebook account in exchange for a money bonus. When participants in a *Direct Tradeoff* Treatment face a choice between a 50 cent bonus and privacy, 64% of participants refuse to share their Facebook profile in exchange for 50 cents. Indeed, when facing a standard price list tool to elicit preferences, the majority of participants in an Elicitation Treatment (who make close to minimum wage) are unwilling to share their Facebook data for \$2.50, and a plurality refuse offers of \$5.00.

However, when the privacy settings are veiled (but revealed costlessly and instantly with the click of a button, as in a moral wiggle room experiment (Dana, 2007), many participants keep themselves in the dark and opt for more money. Participants in a *Veiled Tradeoff* Treatment face a choice between a 50 cent bonus and a 0 cent bonus. They know that one of these bonuses will mean giving out their Facebook profile, and they can click a button to check which option involves a loss in privacy. I find that most people (58%) do not click, and only 40% end up

keeping their Facebook profile private. Hence, people who are willing to pay nearly an hour's worth of wages to stay private are also able to throw caution to the wind, take a 50 cent bonus, and hope for the best.

Importantly, this same avoidance pattern does not hold when participants make a choice between two money bonuses, rather than money versus privacy. In a *Placebo Veiled Tradeoff* Treatment, participants face the exact same experimental interface as in the Veiled Tradeoff group, but where the second column contains a second money bonus. The size of the second money bonus is drawn from the distribution of willingness-to-pay prices from participants in the Elicitation Treatment. When facing this choice, participants in the Placebo Veiled Tradeoff clicked to reveal the second column 66% of the time, a rate significantly different from the reveal rate in the Veiled Tradeoff group.

This paper also presents data on changes in privacy preferences before, during, and after the Cambridge Analytica / Facebook scandal, a major scandal that made privacy issues more salient for many Facebook users. If people are not aware of privacy issues, scandals like this one might disabuse them of this lack of awareness, helping to resolve the inconsistency in people's privacy behavior. By happenstance, an initial round of the experiment was run several weeks before the scandal became public. Once the scandal broke, privacy issues -- and more specifically, privacy issues surrounding Facebook data and third party apps -- dominated the news, appearing on the front page of the New York Times on most days for a month. The experiment was re-run twice with new participants, once at the peak of the scandal and again a month later.

I find that privacy preferences did not change during the scandal, but information avoidance behavior diminished. When facing the Direct Tradeoff treatment, 64% of participants

chose to keep their Facebook profile private instead of getting 50 cents, compared to 67% before the scandal (a slight but statistically insignificant drop). However, participants in the Veiled Tradeoff treatment were more likely to click to learn the privacy setting before making their choice, ultimately resulting in 58% opting for privacy (compared to 40% before the scandal). But this effect was short-lived. Forty days later, 46% opted for privacy over 50 cents -- a proportion statistically indistinct from the pre-scandal level, and significantly lower than the peak-of-the-scandal level.

The results of the experiment make people's inconsistency over privacy choices more mysterious. Until now in the literature, the two dominant explanations of the privacy paradox were revealed preference and ignorance. That is, maybe people give away their data because they value the services they get in return. Alternatively, maybe they do not realize they are giving away their data. In contrast, this experiment finds that the inconsistency persists, even in a setting where participants *know* the exact privacy loss at stake, and where ignorance is unlikely to affect the Direct Tradeoff and Veiled Tradeoff groups differently. At the same time, the experiment also shows that people *are* willing to pay for privacy. Therefore, the inconsistency cannot be written off as mere talk.

The results also cast doubt on current privacy law doctrine in the United States, which relies on giving consumers better notice before they make privacy decisions. Such a policy makes sense if people's privacy inconsistency is explained by revealed preference or ignorance, since either way, better disclosure helps people make better choices. This experiment shows that such a policy will be difficult to execute, because even when, as in this experiment, a privacy disclosure is two words long (“high privacy” vs “low privacy”), many people are willing to avoid the disclosure and give away their data.

## **2. Background**

This section gives a brief background on privacy law and scholarship in three parts. First, it describes how existing privacy law in the United States relies on giving people information about data collection. Second, it describes the research on privacy that has led scholars to conclude that people are inconsistent about privacy choices because they are ignorant or boundedly rational. Third, it shows that information avoidance -- a phenomenon well-documented in other domains -- is an alternative explanation for people's inconsistency.

### **2.1. Privacy law relies on giving people information**

Firms in the United States can legally harvest data from consumers, so long as consumers receive proper notice and agree to the exchange. This framework, known as Notice and Choice, is the standard in United States privacy law (Strahilevitz, 2010). This Notice and Choice model was first outlined in a 1973 report by the U.S. Department of Health, Education and Welfare, and at the time, this legal framework was a departure from how privacy law originally developed. Before the rise in internet commerce and telecommunications, privacy was governed by tort law (Brandeis, 1890; Prosser, 1960; Posner, 1978). So long as it was not the government invading privacy -- in which case constitutional protections would be relevant -- a person could enforce various common law rights to privacy under private causes of action (e.g., a right to seclusion). As private data has become dominated by internet transactions, privacy law has been increasingly governed by contract law principles.<sup>1</sup>

---

<sup>1</sup> There is more stringent regulation for certain consumers and certain industries. Banks send annual privacy notices because of the Gramm-Leach-Bliley Act. Doctors require patients to sign an extra form because of the Health Insurance Portability and Accountability Act. Websites ask users if they are older than 13 -- not 18, not 12, not 16 -- because of the Children's Online Privacy Protection Act. Outside the United States, there is more stringent regulation still. The European Union has started enforcing the General Data Protection Regulation, which imposes stronger consent requirements for data collection, forces firms to delete personal data at a consumer's request, and allows for fines up to 4% of a firm's global revenue.

Since privacy is governed by free choice, it becomes important to understand when and why consumers sell their personal data. As a result, much of the empirical literature on privacy looks at how much consumers value keeping their data private in voluntary transactions.

## **2.2. Privacy preferences are fickle**

The question “how much do people value privacy” has been challenging to answer because people's privacy decisions are fickle. Acquisti (2013) offer people gift cards in exchange for completing a survey. When endowed with a \$10, anonymous gift card, about half of participants chose to keep it rather than exchange it for a \$12, non-anonymous gift card. When endowed with the less private \$12 card, 90% of participants chose to keep it rather than exchange it for the \$10, anonymous card. John (2011) find that people volunteer more sensitive information when asked indirectly, and also when a website seems *less* professional. Similarly, an experiment by legal scholars testing different disclosure techniques finds that people's privacy behavior is not much affected by providing them more and better information about their privacy choices (Ben-Shahar, 2016; *c.f.* Bakos, 2014). Along the same lines, Athey et al (2017) conduct a field experiment where MIT students are given Bitcoin and are invited to start using one of four digital wallets, with varying levels of privacy and convenience. Students' wallet choices were affected by the order in which the wallets were presented, and students' self-reported privacy preferences had no predictive power for their privacy choices. Hence, people's privacy decisions appear inconsistent.

There are two simple explanations for this inconsistency: ignorance and revealed preference.

Under the ignorance explanation, people are unaware of how much data they are emitting, or they struggle to value privacy, because it is abstract or because privacy costs are

inchoate and uncertain, both in scope and timing (Acquisti, 2013). Either way, they do not fully understand what is at stake. As a result, when deciding whether to exchange privacy for something more easily quantifiable, like money or convenience, small frictions may play an outsized role in decision-making. This line of scholarship draws on classic findings from psychology and economics, like the endowment effect and framing effects, to explain people's fickle privacy preferences.

Under the revealed preference explanation, people give up privacy simply because this maximizes their utility. People say they do not like losing privacy, but people also say that they do not like losing \$5. That does not mean it is a paradox if lines of people in a Starbucks happily give away \$5 to a barista each morning -- provided they get a fancy latte in return.

For either explanation -- revealed preference or ignorance -- more information is better. If it's costless, better information will help people make more informed choices in line with their preferences. Or, if people struggle to make consistent choices, better information can help dispel the cognitive biases or lack of awareness that might drive this inconsistency.

### **2.3. Fickle privacy preferences can be explained by bounded rationality *or* by information avoidance**

The privacy literature points to bounded rationality or revealed preference to explain the privacy paradox, but information avoidance can just as easily explain the same pattern. There is a robust literature from psychology and economics on information avoidance (Loewenstein, 2017). While economists typically model information as an intermediate good (Posner, 1978; Stigler, 1961) -- i.e., valuable only because it helps us achieve ends -- scholars in psychology and economics increasingly recognize that people sometimes behave as if information has emotional valence (Oster, 2013). More information is not always better.

Consider a now widely-replicated experiment on moral wiggle room (Dana, 2007), which is the basis for the experimental design used in this paper. In the experiment, a participant has to choose payoffs for herself (“me”) and a partner that she does not meet (“my partner”). In a baseline condition, she chooses between two options: \$6 for me and \$1 for my partner, or \$5 for me and \$5 for my partner. Most people pick the second option. A treatment group faces a slightly modified choice: \$6 for me and \$X for my partner, or \$5 for me and \$Y for my partner. In this case, either X is 1 and Y is 5 (as in the baseline group), or X is 5 and Y is 1. The person can costlessly click to find out the values of X and Y.

Consider what a typical economic model would predict. If, in the baseline experiment, I preferred \$5 and \$5 over \$6 and \$1, and this is a strong preference, then I should click to find out the value of X and Y. Either I find out that I am in the baseline case, in which case I can choose \$5 and \$5 again, or I will find out that I am in the easier case and choose \$6 and \$5.

But this is not how people act in the experiment. Instead, people avoid learning the values of X and Y and pick the \$6 for me, \$X for my partner option. They exploit the wiggle room to act selfishly. Other experiments on altruism, lab- and field-based, find similar results (Exley, 2016; Malmendier, 2012; List, 2002).

This pattern of behavior is important across disparate domains. In health, one study found that 27% of intravenous drug users at risk of HIV who got tested did not return to the clinic to see their results (Sullivan, 2004), even though knowing one's HIV positive status can lengthen one's life. In family planning, twenty states have laws requiring women to see a picture of the fetus before getting an abortion (Guttmacher Institute, 2018). Presumably, women know what a fetus looks like, so the law was not passed because the increased information of the fetus's



appearance will lead to more informed choices. In sum, people avoid information that upsets them, even if in theory a utility-maximizing agent would never reject free information.

Given the central focus in privacy law on giving consumers better, cheaper information, and given the psychology and economics literature on how people avoid information, this paper focuses on testing an important open question: do people engage in information avoidance when making privacy decisions?

### **3. Experimental Design**

I conduct an experiment to test for information avoidance in privacy decisions. Participants are randomized to one of two treatments: a Direct Tradeoff Treatment and a Veiled Tradeoff Treatment. This section first discusses the overall timeline of the experiment, then describes the two treatments in detail.<sup>2</sup>

795 participants were recruited on Amazon Mechanical Turk to take a short survey about health and financial status.<sup>3</sup> All participants were informed that before doing the survey, they would make decisions about the size of a bonus payment, to be received upon completion, and the privacy settings of the survey.<sup>4</sup>

After recruitment, the timeline of the experiment consists of three stages: instructions and practice, privacy settings, and a survey.<sup>5</sup> First, participants were shown an initial introductory screen that gives an overview of their participation. Participants were told that they would take a

---

<sup>2</sup> The experiment was approved by Harvard's Committee on the Use of Human Subjects as protocol IRB18-0061. The experiment was pre-registered on AsPredicted under the title "Information Avoidance and Internet Privacy" (#16702)

<sup>3</sup> Research increasingly suggests that, for the purpose of social science experiments, Mechanical Turk users are a reliable sample. Irvine (2018) replicates three experiments using in-person labs, national online platforms, and Mechanical Turk, and finds that the results are constant across samples. The key difference was that that Mechanical Turk users were significantly more attentive than the other samples. See also Hoffman (2017), which replicates an experiment on Mechanical Turk, on college students in a physical lab, and college students in an online setting.

<sup>4</sup> Median hourly wages for workers was \$14.96 (based on a median payment of \$1.52 for a median completion of 8 minutes 6 seconds). The experiment was conducted on November 20, 2018.

<sup>5</sup> Appendix A presents the entire experimental instructions.

survey, but while everyone would take the same exact survey, each participant would be given a choice between two privacy options. They could opt for high privacy, in which case their survey answers would be anonymous. Or, they could opt instead for low privacy, in which case they would click a “Log In with Facebook” button at the top of the survey. This meant that the survey-taker would see, in addition to the participant's survey answers, her public Facebook profile (including profile picture, name, and gender) and her email address. Participants who chose low privacy would not be allowed to finish the survey until they logged in. Participants then completed two short practice rounds which looked identical to the privacy settings task.

After the instructions stage, participants chose their privacy settings. After completing the privacy settings stage, participants completed the survey stage.

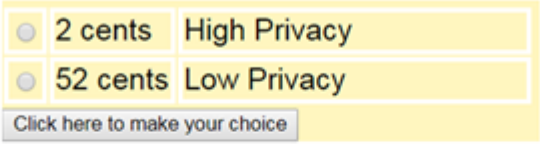
The privacy measure in the experiment -- whether to share Facebook information -- has three advantages: it is a real decision, it is a realistic one, and it is an important one. First, participants who give up their privacy in this experiment must actually give over their profile data, so the choice is not a hypothetical one. Nor is it a behavior that can be faked: unlike other privacy experiments, which measure privacy as a person's willingness to answer an intrusive question, a participant in this experiment cannot pretend to give up privacy without actually giving anything up.<sup>6</sup> Second, the decision is a realistic one. The “Log In with Facebook” button is a ubiquitous part of the internet -- many websites allow people to log in with their Facebook (or Google) account rather than with the website itself. Hence, it is a choice people routinely make: should I engage in online activity in a way that is linked to my Facebook profile or not?

---

<sup>6</sup> Even if participants have a fake account they can use -- Facebook works hard to limit such behavior, but is not 100% successful -- even handing over a fake account involves some cost. Doing so means the experimenter can link a fake Facebook account to a Mechanical Turk account (and the answers in the survey), which makes the fake account less effective.

Third, the decision has important public policy implications, as suggested by the Cambridge Analytica scandal.

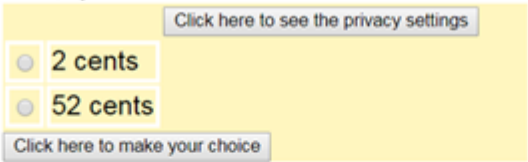
Each person was randomized into one of two treatments during the privacy settings stage: the Direct Tradeoff Treatment and the Veiled Tradeoff Treatment. Figure 1.1 shows the exact format of the privacy choice made in each of the treatments.

- On the next screen, you'll see a table like this:  


|                                |              |
|--------------------------------|--------------|
| <input type="radio"/> 2 cents  | High Privacy |
| <input type="radio"/> 52 cents | Low Privacy  |

Click here to make your choice
- The first column shows how big your bonus will be.
- The second column shows your privacy settings. “High Privacy” means doing the survey anonymously. “Low Privacy” means doing it after logging in through Facebook.
- Sometimes, the top row will be the high privacy option. Sometimes it will be the low privacy option. It can be either, with a 50/50 chance.

- On the next screen, you'll see a table like this:  


|                                |  |
|--------------------------------|--|
| <input type="radio"/> 2 cents  |  |
| <input type="radio"/> 52 cents |  |

Click here to see the privacy settings

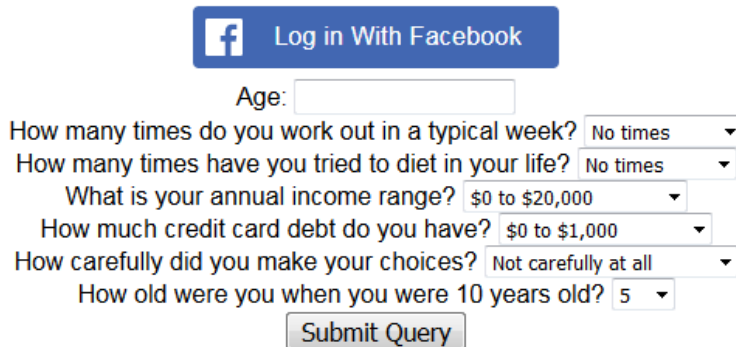
Click here to make your choice
- The first column shows how big your bonus will be.
- The second column – which you have to click to see – shows your privacy settings. “High Privacy” means doing the survey anonymously. “Low Privacy” means doing it after logging in through Facebook.
- Sometimes, the top row will be the high privacy option. Sometimes it will be the low privacy option. It can be either, with a 50/50 chance. You have to click to make sure.

Figure 1.1: This figure shows the instructions page for each of the two treatments. The Direct Tradeoff Treatment group was shown the instructions in the top panel. The Veiled Tradeoff Treatment group was shown the instructions in the bottom panel.

In the Direct Tradeoff Treatment, participants only made one decision: a direct choice between a \$0.02 bonus and Privacy Option A or a \$0.52 bonus and Privacy Option B. The privacy options were randomized so that half the time, participants faced a degenerate choice between (more money, more privacy) and (less money, less privacy). The other half of the time, participants faced a true tradeoff between money and privacy.

In the Veiled Tradeoff Treatment, participants faced the same decision as in the Direct Tradeoff Treatment, but the privacy setting was initially hidden. Participants had to click to reveal the column describing the privacy settings, and there was a 50% chance that the higher money bonus would mean losing their anonymity.<sup>7</sup>

After completing the privacy task, all participants completed a nine-question survey, shown in Figure 1.2.



The image shows a survey form with a blue button at the top that says "Log in With Facebook" with a Facebook logo. Below the button are several questions with input fields or dropdown menus:

- Age:
- How many times do you work out in a typical week? No times (dropdown)
- How many times have you tried to diet in your life? No times (dropdown)
- What is your annual income range? \$0 to \$20,000 (dropdown)
- How much credit card debt do you have? \$0 to \$1,000 (dropdown)
- How carefully did you make your choices? Not carefully at all (dropdown)
- How old were you when you were 10 years old? 5 (dropdown)

At the bottom of the form is a button labeled "Submit Query".

Figure 1.2: After making their privacy choices, all participants completed the survey above. Those participants who opted for the anonymous survey were not shown the Facebook login button. Those that opted for the low privacy setting saw the login button, as in the picture above.

Five questions covered demographics, health, and financial topics. These questions asked about the person's age, the number of times they exercise in a week, the number of times they

---

<sup>7</sup> Note that for both groups, there was a 50% chance of facing a degenerate choice between (more money, more privacy) and (less money, less privacy). These decisions cannot tell us about how much a person values privacy, so they are omitted from the main analyses below. The resulting sample size is 535 participants: 117 in the Direct Tradeoff, 130 in the Veiled Tradeoff, 164 in the Placebo Veiled Tradeoff (described below), and 124 in the Elicitation Treatment (described below).

have attempted to diet in their life, their annual income, and their credit card debt. The survey also included two questions to check comprehension. One asked “How old were you when you were 10 years old” with a dropdown menu with several options, including 10. Another directly asked “How carefully did you make your choices?” with three options: “Not carefully at all”, “I thought about it a little”, and “I was very careful”. Two questions asked whether participants had a Facebook profile and how often they used Facebook. After submitting the survey, participants were finished.

The user interface for the experiment was coded using HTML and Javascript, which ensured that the “reveal button” would work instantaneously -- without a page refresh. When a user clicked the reveal button, Javascript code changed the visibility setting of the hidden column from hidden to visible. The hidden column would therefore become visible immediately. The users' choices and data were sent to a MySQL database using PHP code. All code is available on request from the author and includes survey instructions, experimental module coding, and the raw data.<sup>8</sup>

### **3.1. Placebo Test**

Any difference between the Direct Tradeoff and Veiled Tradeoff groups might be driven by clicking costs, rather than information avoidance. Suppose many people value privacy at 51 cents, but the “click to reveal” button imposes a few cents of effort costs. Then we would observe a treatment effect, but because participants rationally conclude that it's not worth spending a few cents of effort for a 1 cent gain.

I test this alternative explanation in two ways. First, I use an Elicitation Treatment to gather the full distribution of willingness-to-pay (WTP) prices for privacy. In the Elicitation

---

<sup>8</sup> Contact the author for the ZIP file: [dsvirsky@hbs.edu](mailto:dsvirsky@hbs.edu)

Treatment, instead of making just one choice between privacy and 50 cents, participants made 10 choices, with the money bonus varying between 25 cents and \$5.00. Participants were told that one of their choices would be enforced. This is a standard technique in applied microeconomics to elicit a WTP price, in this case for staying anonymous. Participants faced a table as in Figure 1.1, in which they chose between two rows of a table. The top row meant a \$0.02 bonus and “High Privacy”, and the bottom row meant a \$X.YY bonus and “Low Privacy”, with \$X.YY ranging from \$0.27 to \$5.02. Hence, if someone opted to stay anonymous when offered \$0.50, \$1.00, and \$1.50, but not at \$2.00, then we can infer that her WTP for staying anonymous is between \$1.50 and \$2.00.

Second, I conduct a Placebo Veiled Tradeoff treatment. This treatment is identical to the Veiled Tradeoff treatment, but instead of making a choice between one money bonus and privacy, participants make a choice between one money bonus and a second money bonus. The first money bonus is 50 cents, as in the main experiment, but the second money bonus is randomly drawn from the distribution of WTP prices from the Elicitation Treatment. If clicking costs alone are driving results in the main experiment, where people have some distribution of WTP prices for privacy, then we would observe the same size treatment effect if the second column is instead a money bonus drawn from the same distribution of WTP prices.

### **3.2. Privacy Preferences Before, During, and After the Cambridge Analytica / Facebook Scandal**

On March 18, 2018, The Guardian first reported that Cambridge Analytica, a political consulting firm, had harvested data from nearly 90 million Facebook accounts in order to help conservative political candidates. Most of the data was obtained without consent, and the report quickly escalated into a public scandal. Cambridge Analytica largely relied on Mechanical Turk

to construct its illicit dataset. Mechanical Turk users were invited to share their Facebook data in exchange for monetary bonuses between \$2 and \$4, but in addition, the users gave permission to Cambridge Analytica (under false pretenses) to access their friends' profile data as well. The option to share friends' data was discontinued in 2016.

The specific nature of the scandal could not have been better-suited to the dependent variable for privacy used in this experiment. Specifically, the scandal dealt with people's willingness to share their Facebook data as part of an unrelated survey, which is precisely the dependent variable measured in this paper. Further, Cambridge Analytica targeted Mechanical Turk users, so the experiment in this paper was run on the same sample of people targeted in the scandal -- though most likely not the exact same people, given natural turnover rates in Mechanical Turk's worker base.

The experiment was run three times, and the timing was chosen to measure whether privacy preferences changed during and after the scandal. The pilot round of the experiment was initially run on February 23, 2018 -- 23 days before the scandal broke. A second round was conducted 11 days after the scandal became public. A third round was conducted 41 days later.

Figure 1.3 uses Google trends data to show how often people searched for the phrase "Facebook privacy settings". The graph shows a spike in such searches in the immediate aftermath of the scandal, coinciding with the second round of the experiment. This spike in search interest diminished by the time the third round was conducted.

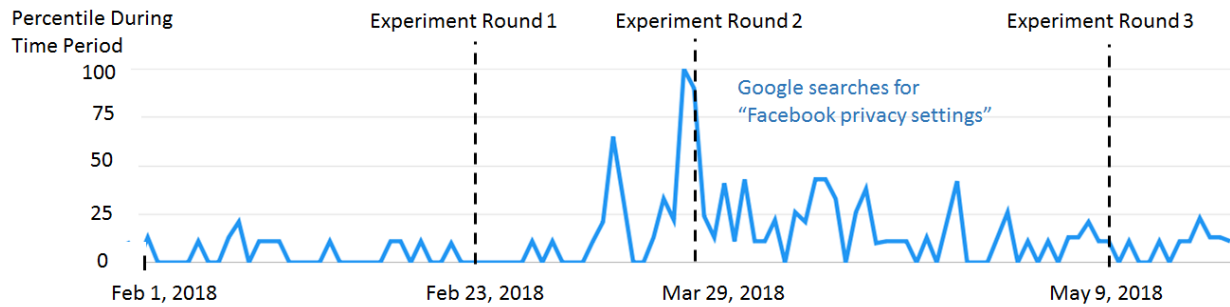


Figure 1.3: This figure shows the relative volume of Google searches for the phrase “Facebook privacy settings” over time, as well as the timing of the initial three rounds of the experiment. These three rounds, all identical in design, are used to measure changes in privacy valuations and information avoidance behavior during the Facebook Cambridge Analytica Scandal.

The main results presented in this paper are from an experimental round run on November 20, 2018, whereas the Facebook results are from three earlier rounds of the experiment. These initial three rounds did not include an Elicitation Treatment, only a Direct Tradeoff and Veiled Tradeoff treatment. Importantly, the three initial rounds were all identical to each other, which ensures that comparisons across these three rounds are valid.

#### 4. Results

Table 1.1 presents summary statistics on the survey answers, as well as a balance check. Nearly all -- 94.6% -- participants reported having a Facebook account. This is important, as it is not clear how a person without a Facebook account would make a valuation decision in this experiment (though the balance check confirms that, however this would affect results, the lack of a Facebook account was similar across treatments). All analyses are substantively unchanged whether these participants are excluded or included, but in the data below, they are included. Across participants, Facebook use was common. The median participant reported using Facebook 4 or more times per week.



Table 1.1: Summary Statistics

|  | Direct Tradeoff | Veiled Tradeoff | P-Value |
|--|-----------------|-----------------|---------|
| Age (Years)                                  | 31.0<br>(7.3)   | 32.5<br>(8.6)   | 0.13    |
| Diet Attempts in Lifetime (0 - 4+)           | 2.3<br>(1.6)    | 2.3<br>(1.6)    | 0.93    |
| Exercise Workouts in a Typical Week (0 - 4+) | 2.4<br>(1.4)    | 2.3<br>(1.4)    | 0.49    |
| Annual Income (0 - 4)                        | 1.3<br>(1.1)    | 1.1<br>(1.2)    | 0.14    |
| Credit Card Debt (0 - 4)                     | 0.73<br>(1.1)   | 0.80<br>(1.1)   | 0.66    |
| Has Facebook (0,1)                           | 0.92<br>(0.28)  | 0.95<br>(0.21)  | 0.21    |
| Weekly Facebook Use (0 - 4+)                 | 2.8<br>(1.5)    | 3.0<br>(1.4)    | 0.31    |

Table 1.1: Summary statistics for the Direct Tradeoff and Veiled Tradeoff groups. Standard deviation reported in parenthesis. Each statistic is taken from the participants' survey answers. Income and credit card debt variables are categorical. Each category from 0 to 4 represents a different income or debt range. Diet attempts, Exercise and Facebook use can be 0, 1, 2, 3 or “4 or more.” Reported p values taken from t-tests comparing the means of the two groups.

The analyses below are restricted to participants who answered both the privacy valuation task and the survey, but attrition from the study may be of substantive interest in its own right, for example if people drop out of the study when they see that they have to share Facebook information. Attrition was quite low. In the Direct Tradeoff and Veiled Tradeoff treatments, attrition (defined as people who read the instructions but quit before the survey round) was 5% and 2% respectively.

#### 4.1. Results: Direct Tradeoff Treatment vs Veiled Tradeoff Treatment

I find a treatment effect from putting a costless veil on privacy settings. 64% of people in the Direct Tradeoff Treatment refuse to sell their Facebook data for 50 cents.<sup>9</sup> In contrast, in the

<sup>9</sup> This is in line with the results from the Elicitation Treatment group, described below. In that group, 59% rejected an offer of 50 cents to share their Facebook profile, a slightly lower but statistically insignificant difference.

Veiled Tradeoff Treatment, when the privacy consequences of their actions are initially hidden, only 40% refuse to sell their Facebook data for 50 cents. A majority in the Veiled Tradeoff Treatment (58%) chose *not* to look at the privacy setting before deciding to take the 50 cents.

Figure 1.4 result shows the proportion of participants who remained anonymous in the Direct Tradeoff Treatment and the Veiled Tradeoff Treatment. Figure 1.5 breaks down participants' decisions in both treatments, including their privacy choice as well as their decision whether to click. Table 1.2 reports various regressions where the unit of observation is an individual, the dependent variable is whether the participant ended up remaining anonymous, and the independent variable is an indicator variable for being in the Veiled Tradeoff Treatment.

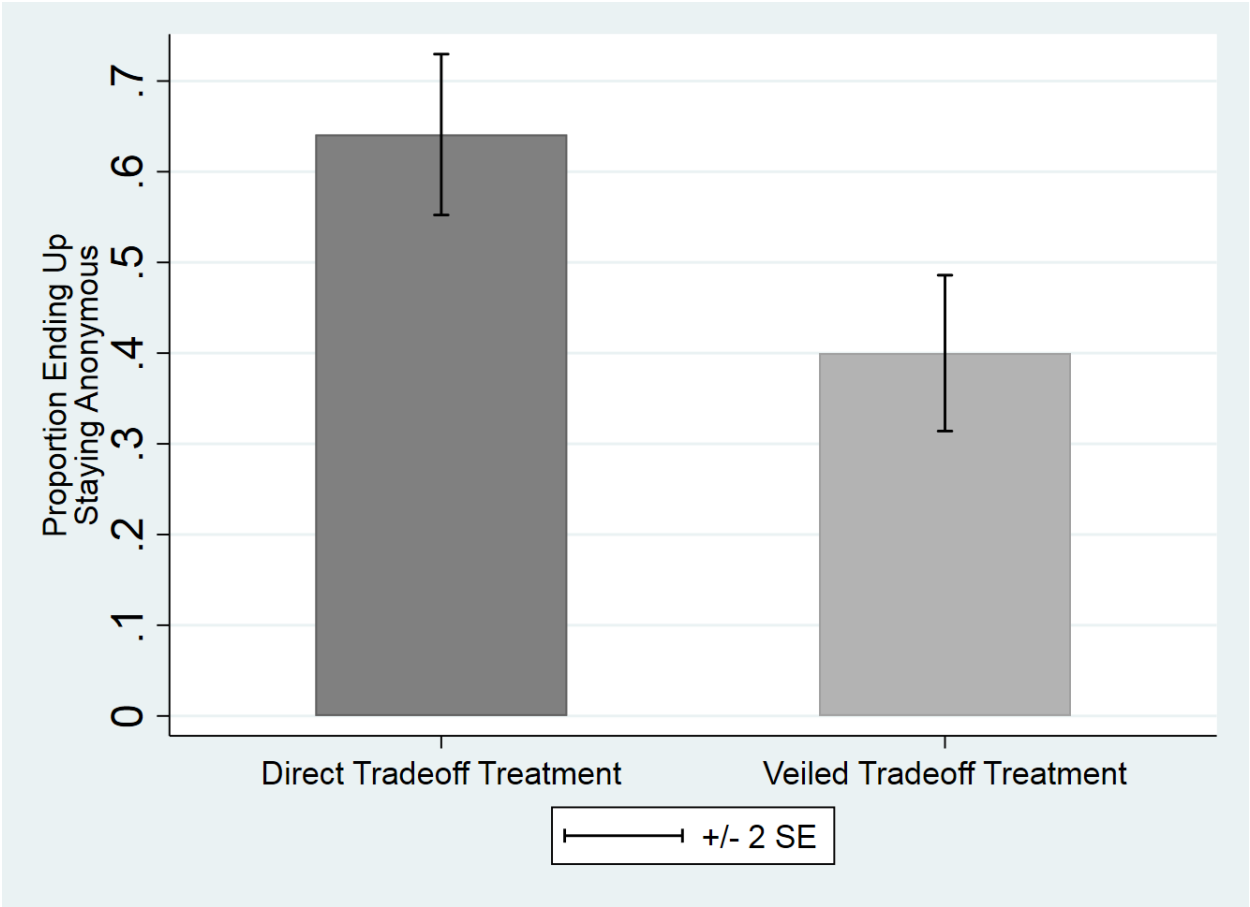


Figure 1.4: This figure shows the proportion of participants who ended up remaining anonymous instead of sharing their Facebook profile for 50 cents, for the Direct Tradeoff Treatment (N = 117) and the Veiled Tradeoff Treatment (N = 130). These results exclude all participants who, by

randomization, faced a degenerate tradeoff of 50 cents and high privacy vs 0 cents and low privacy. Therefore, for the Veiled Tradeoff Treatment, anyone who chose the higher money option is counted as having chosen 50 cents over anonymity, regardless of whether they clicked to reveal the privacy setting before making their decision.

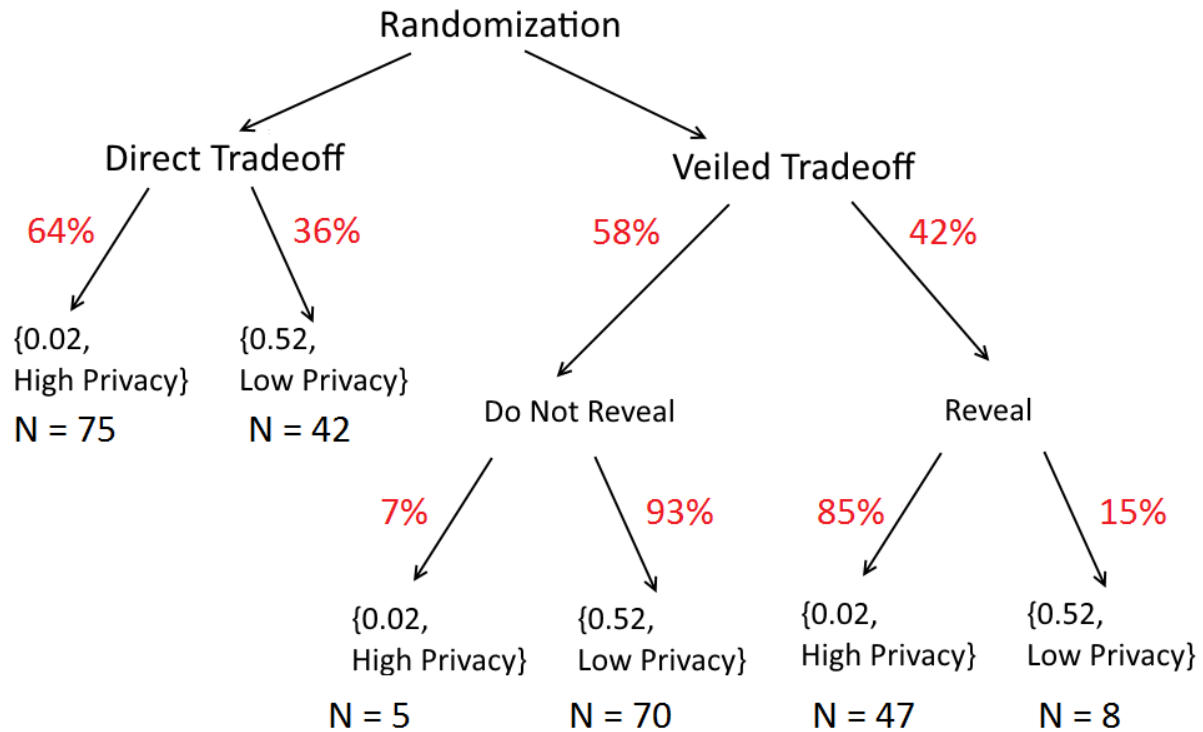


Figure 1.5: This figure shows the decisions made by participants across the Direct Tradeoff Treatment and Veiled Tradeoff Treatment. This figure excludes participants who were randomized into a degenerate choice between more money *and* high privacy vs less money and low privacy. In the Direct Tradeoff Treatment, participants made a choice between {0.02, High Privacy} versus {0.52, Low Privacy}. In the Veiled Tradeoff Treatment, participants first decide whether to reveal or not to reveal. If they do not reveal, then they choose between {0.02, Privacy Option A} and {0.52, Privacy Option B}. If they do reveal, then they face the same choice as in the Direct Tradeoff Treatment. Because I exclude all participants who face a degenerate choice, the lower monetary bonus always corresponds to high privacy, though participants in the Veiled Tradeoff Treatment who do not click to reveal cannot be certain of this, and only know that there is a 50% chance that low money corresponds to high privacy and a 50% chance that low money corresponds to low privacy.

Table 1.2: Privacy Decisions in Veiled and Direct Tradeoff Groups, with Robustness Checks

| Sample                        | (1)<br>Full Sample | (2)<br>Full Sample | (3) Passed<br>Comprehension Check | (4) Answered<br>Carefully | (5) Intersection of<br>(3) and (4) |
|-------------------------------|--------------------|--------------------|-----------------------------------|---------------------------|------------------------------------|
| Veiled Tradeoff Group         | -0.24***<br>(0.06) | -0.23***<br>(0.07) | -0.21**<br>(0.07)                 | -0.22***<br>(0.06)        | -0.19**<br>(0.07)                  |
| Age                           |                    | 0.01**<br>(0.004)  |                                   |                           |                                    |
| Diet Attempts<br>in Lifetime  |                    | -0.01<br>(0.02)    |                                   |                           |                                    |
| Exercise in a<br>Typical Week |                    | -0.04<br>(0.02)    |                                   |                           |                                    |
| Annual Income                 |                    | 0.02<br>(0.03)     |                                   |                           |                                    |
| Credit Card Debt              |                    | 0.03<br>(0.03)     |                                   |                           |                                    |
| Constant                      | 0.64***<br>(0.05)  | 0.34*<br>(0.17)    | 0.63***<br>(0.05)                 | 0.63***<br>(0.05)         | 0.62***<br>(0.05)                  |
| Observations                  | 247                | 213                | 207                               | 239                       | 201                                |
| Adjusted $R^2$                | 0.05               | 0.07               | 0.04                              | 0.04                      | 0.03                               |

Table 1.2: Table reports a linear probability regression of a binary variable for whether the participant ended up staying anonymous on a binary variable for whether the participant was in the Veiled Tradeoff Treatment. Omitted group is the Direct Tradeoff Treatment. Robust standard errors in parentheses. \* $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

The treatment effect is robust even if we exclude participants who failed comprehension and attention checks. During the survey (and after completing their privacy choices), one question asked “How old were you when you were 10?” with several options in a dropdown menu. Roughly 84% of participants correctly answered. Another question in the survey asked “How carefully did you make your choices?”, with three options: not carefully, a little carefully, and very carefully. Roughly 75% of participants said they answered the questions “very carefully”, 21% said “a little carefully” and 3% said not at all carefully. Note that by default, “Not Carefully At All” was selected. The main results are substantively unchanged if we exclude participants who did not pay very careful attention or who answered the comprehension question wrong.

Another robustness concern is confusion -- did participants in the Veiled Tradeoff Treatment mistakenly assume that a low bonus meant they would keep their privacy? That is, participants in the Veiled Tradeoff Treatment could have made (incorrect) guesses about the privacy settings, even though the instructions explicitly told them that the privacy settings were randomized. For example, a person could assume that the lower monetary payoff always meant higher privacy. In that case, we would expect that people would choose to never click to reveal the privacy setting but then nonetheless choose the lower payoff. Such behavior occurred in 4% of participants in the Veiled Tradeoff treatment. The results discussed here categorize these participants as having chosen privacy over 50 cents, but the results do not change if these participants are instead dropped.

Table 1.2 reports the results of these robustness checks. Columns 2 - 5 report the result of the main regression, described above, but using different samples. Column 2 includes controls for survey answers, while columns 3 - 5 exclude participants based on comprehension, attention, and confusion (defined as opting for less money without clicking to reveal the privacy setting). The main results hold throughout.

#### **4.2. Results: Placebo Veiled Tradeoff and Elicitation Treatment**

The results from the Elicitation Treatment and Placebo Veiled Tradeoff give strong evidence that clicking costs are not driving the treatment effect in the main experiment. An alternative explanation of the results is that clicking to reveal the privacy settings is costly. It is possible that many participants value privacy at only slightly more than 50 cents, so when faced with the “click to reveal” button, they rationally decide that the costs of clicking and deciding are not worth the small gain in utility of potentially getting privacy over money.

One way to rule this is out is by directly eliciting people's WTP price for sharing their data, and in doing so, I find that the majority of people value privacy at \$2.50 or above. Table 1.3 shows people's WTP for staying anonymous in the Elicitation Treatment. Each row shows the proportion of participants who switched from High Privacy to Low Privacy at the prices offered. The results show that a plurality of participants -- 41.5% -- refuse to share their Facebook profile at all prices, even up to \$5.00. Note that the average hourly wage on Mechanical Turk is roughly \$5 per hour (Hara, 2018), so these participants would rather spend an hour of time completing mundane computer tasks than share their public Facebook profile with a survey-taker. Nonetheless, the second most-common WTP price was at the lower end, with 20.8% choosing to sell their Facebook profile at 25 cents. The remaining 38% evinced a WTP between 25 cents and \$5.00.<sup>10</sup>

Table 1.3: Distribution of Willingness-to-Pay Prices, Elicitation Treatment

| WTP    | Column % | Cumulative % |
|--------|----------|--------------|
| \$0.25 | 20.8     | 20.8         |
| \$0.50 | 20.8     | 41.5         |
| \$1.00 | 1.9      | 43.4         |
| \$2.00 | 1.9      | 45.3         |
| \$3.00 | 3.8      | 49.1         |
| \$4.00 | 9.4      | 58.5         |
| \$5.00 | 41.5     | 100.0        |
| Total  | 100.0    |              |

N = 106

Table 1.3: This table presents the breakdown of Willingness-to-Pay (“WTP”) prices in the Elicitation Treatment. Participants faced 10 binary decisions where they could sell their

<sup>10</sup> Irrational behavior, defined as having multiple switching points, was rare. It is hard to interpret someone giving up her privacy for 50 cents but not for \$1.00, assuming that she also values more money over less money. In the Elicitation Treatment, 84% gave rational answers in the sense of having at most one switching point. This is a relatively low level of multiple switch behavior compared to other experiments that use multiple price lists, which typically find levels of multiple switch behavior ranging from 10% to as high as 50% (Andreoni, 2012; Meier, 2016). This finding also suggests that Mechanical Turk workers evinced similar levels of this type of irrationality when compared to college students and people with moderate incomes in tax filing centers, among other samples. In calculating the distribution of WTP prices, I exclude participants with multiple switches, but the results are similar if I instead include them and define their switching point as either the lowest switch, the highest switch, or the average of the two.

Facebook data for \$X.XX, with \$X.XX ranging from \$0.25 to \$5.00. Participants were informed that one of their 10 decisions would be randomly selected and enforced. From these decisions, a WTP price is calculated by finding the switching point at which a person is willing to begin selling her data. People with multiple switching points are omitted (18 out of 124 participants had multiple switching points). People who refused to sell data at all prices are categorized as having a WTP of \$5.00.

Can the main experimental findings be explained by simple clicking costs? For example, if a user has a WTP for privacy of 50.1 cents, then it might not make sense to take a few seconds to reveal the privacy settings, even if she would have opted for privacy in the direct tradeoff treatment. Using the results of the Elicitation Treatment, I can say with more precision how high clicking costs would have to be to support such an explanation. Appendix B presents a more detailed mathematical approach to this question. It demonstrates three key points. First, anyone who values privacy at less than 50 cents should *never* click to reveal the privacy settings. Second, if clicking costs are zero, anyone who values privacy at more than 50 cents should *always* click to reveal the privacy settings. Third, and most relevant here, clicking costs would have to be nearly \$2.00 to explain the treatment effect in this experiment. I consider this unlikely in this context, especially given that the median participant clicks their mouse 31 during the experiment times and is paid \$1.52 for her participation. If clicking to reveal the two-word privacy settings really imposed a cost of \$2.00, participants would be making a massive mistake by doing this experiment and finishing it.

The results of the Placebo Veiled Tradeoff give more direct evidence that the results are not driven by clicking costs. Recall that in the Placebo Veiled Tradeoff, participants chose between two money bonuses, with the value of the second money bonus drawn from the distribution of WTP prices in the Elicitation Treatment. Participants knew the size of the second bonus and had to click to reveal which row the bonus was in. Among this group, the proportion of participants clicking to reveal the second column was 0.66. This is higher than the click rate of

0.42 in the main experiment, when participants chose between money and privacy, and the difference is statistically significant (Fisher's exact  $p < 0.001$ ). Table 1.4 shows the click proportion in the Placebo Veiled Tradeoff group, broken down by the size of the second money bonus. These results suggest that people are capable of clicking to reveal the second bonus, and do so in a roughly rational way, when money is at stake instead of privacy.

Table 1.4: How Many People Clicked To Reveal in Placebo Veiled Tradeoff Treatment?

| Size of Second Money Bonus | Did Not Click to Reveal | Did Click to Reveal |
|----------------------------|-------------------------|---------------------|
| \$0.25                     | 39                      | 29                  |
| \$0.50                     | 2                       | 1                   |
| \$1.00                     | 1                       | 2                   |
| \$2.00                     | 2                       | 8                   |
| \$3.00                     | 4                       | 8                   |
| \$4.00                     | 0                       | 9                   |
| \$5.00                     | 8                       | 51                  |

N = 164

Table 1.4: This table shows the clicking behavior for participants in the Placebo Veiled Tradeoff treatment, by size of the second money bonus. The table shows how many participants clicked to reveal the size of the second bonus before deciding which row to choose, and how many did not click to reveal before deciding.

In sum, the results of the Elicitation Treatment and Placebo Veiled Tradeoff Treatment suggest that the findings of the main experiment are not driven by clicking costs or confusion about the experimental design.

### 4.3. Privacy Preferences Before, During, and After the Facebook Cambridge Analytica Scandal

Roughly a month after a pilot round of the experiment was run, there was a controversial privacy scandal that directly involved people's willingness to share their Facebook data with third parties. A second and third round of the experiment were therefore run, one in the immediate aftermath of the scandal, and another roughly one month after.



There is no evidence that the sample of participants was observably different across time. Importantly, any changes we see are not necessarily attributable to the scandal, nor is the direction of any effect obvious ex ante. The experiment is limited in the sense that results could be driven by changes in the underlying sample of participants, or trends that affect people's WTP for keeping their Facebook profile private from a third party but that were unrelated to the Facebook scandal. To get a sense of these issues, Table 1.5 presents a balance check to see whether the three samples of participants are significantly different in any of the survey responses. I find balance across all three groups, suggesting that in terms of reported age, credit card debt, income, and exercise patterns, the sample did not measurably change before, during, and after the scandal.

Table 1.5: Summary Statistics Before, During, and After Facebook Scandal

|                                     | Before Scandal   | During Scandal   | After Scandal    | P-Value |
|-------------------------------------|------------------|------------------|------------------|---------|
| Age                                 | 33.54<br>(10.67) | 32.40<br>(9.711) | 33.06<br>(8.356) | 0.12    |
| Diet Attempts in Lifetime           | 2.32<br>(1.58)   | 2.26<br>(1.49)   | 2.31<br>(1.56)   | 0.85    |
| Exercise Workouts in a Typical Week | 2.48<br>(1.33)   | 2.36<br>(1.34)   | 2.40<br>(1.44)   | 0.55    |
| Annual Income                       | 1.33<br>(1.11)   | 1.33<br>(1.15)   | 1.23<br>(1.12)   | 0.20    |
| Credit Card Debt                    | 0.64<br>(0.99)   | 0.71<br>(0.95)   | 0.64<br>(1.01)   | 0.98    |

Table 1.5: Summary statistics for all participants, broken down by whether the sample was from before, during, or after the Facebook scandal. To calculate p-value for a row, the variable for the survey response was regressed on indicator variables for two of the three treatments. The p-value reported is the p-value for the F-test, or the joint hypothesis that all the coefficients are insignificant.

At the height of the Facebook / Cambridge Analytica scandal, people's behavior in the Direct Tradeoff Treatment was unchanged. Before the scandal, 66% opted for privacy over 50

cents in the Direct Tradeoff Treatment. At the height of the scandal, this number was 64%, and one month later, the proportion was 63%. None of these changes were statistically significant.

However, the treatment became less effective. Before the scandal, the Veiled Tradeoff Treatment caused a 26 percentage point drop ( $p < 0.001$ ) in the proportion of people opting to keep their Facebook profile private. At the height of the scandal, the Veiled Tradeoff Treatment caused a 9 percentage point drop ( $p = 0.06$ ). One month after the scandal, the treatment was effective again, causing a 17 percentage point drop ( $p = 0.003$ ). The treatment effect at the height of the scandal was significantly different from the treatment effects before ( $p = 0.01$ ) and after the scandal ( $p = 0.03$ ).

Figure 1.6 shows the proportion of people who chose to keep their Facebook data private during the survey instead of getting a fifty cent bonus, by treatment and across the three experiment dates.

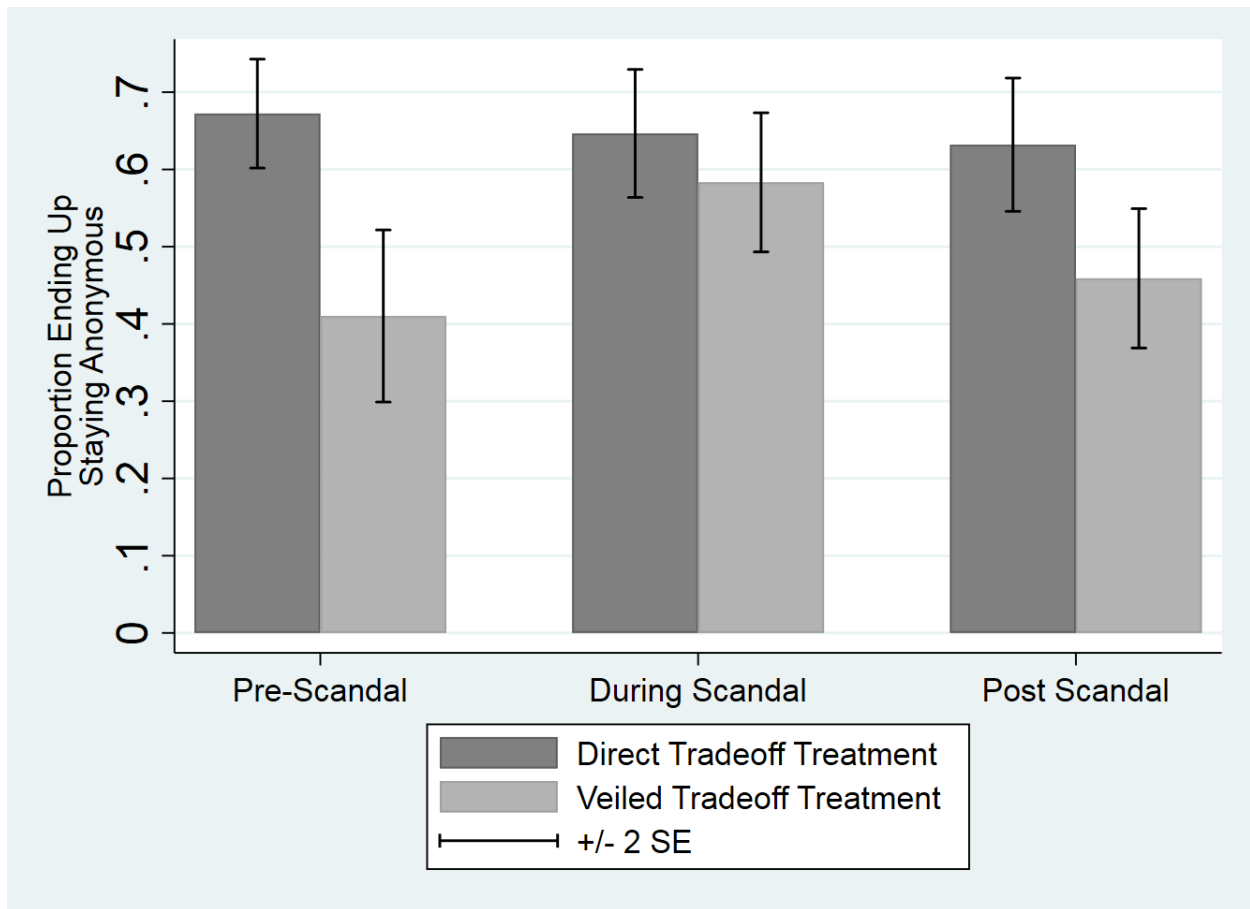


Figure 1.6: This figure shows the proportion of people in the Direct Tradeoff and Veiled Tradeoff treatments who ended up staying anonymous instead of getting a 50 cent bonus, across experiment dates. The Facebook scandal became public on March 18, 2018. The first round of the experiment occurred on February 23. The second round occurred on March 29. The third round occurred on May 9. Error bars are +/- two standard errors.

Table 1.6 presents regression results and robustness checks. The regression specification is as follows, letting  $p$  be an indicator variable for whether an individual ended up remaining anonymous,  $T$  be an indicator variable for whether the participant was in the Veiled Tradeoff Treatment,  $FB$  be an indicator for whether the experiment date occurred shortly after the Facebook scandal, and  $Post$  be an indicator for whether the experiment occurred forty days after the scandal.

$$p = \beta_0 + \beta_1 \cdot T + \beta_2 \cdot FB + \beta_3 \cdot FB \cdot T + \beta_4 \cdot Post + \beta_5 \cdot Post \cdot T + \epsilon_i$$

In the regression,  $\beta_1$  measures the treatment effect before the scandal,  $\beta_2$  measures the change in privacy preferences in the Direct Tradeoff Treatment group at the height of the Facebook scandal,  $\beta_3$  measures the change in the treatment effect at the height of the Facebook scandal,  $\beta_4$  measures the change in privacy preferences in the Direct Tradeoff Treatment group after the scandal, and  $\beta_5$  measures the change in the treatment effect after the scandal. Column 1 includes the entire sample. Column 2 excludes participants who failed the comprehension check. Column 3 excludes participants who reported not answering carefully. Column 4 excludes participants who did not click to reveal the privacy setting but chose the lower money option.

Table 1.6: Treatment Effect Before, During, and After Facebook / Cambridge Analytica Scandal

|  | (1)                | (2)                        | (3)                | (4)                                    |
|--|--------------------|----------------------------|--------------------|--|
| Sample   | Full Sample        | Passed Comprehension Check | Answered Carefully | Excludes 'Didn't Click, Chose 0 cents' |
| Privacy Setting Hidden                                 | -0.26***<br>(0.06) | -0.21**<br>(0.08)          | -0.19**<br>(0.07)  | -0.30***<br>(0.07)                     |
| During Facebook Scandal                                | -0.03<br>(0.05)    | 0.00<br>(0.06)             | -0.00<br>(0.06)    | -0.03<br>(0.05)                        |
| During Facebook Scandal<br>* Veiled Tradeoff Treatment | 0.20*<br>(0.08)    | 0.14<br>(0.11)             | 0.15<br>(0.10)     | 0.21*<br>(0.10)                        |
| Post-Facebook Scandal                                  | -0.04<br>(0.05)    | -0.01<br>(0.06)            | -0.00<br>(0.06)    | -0.04<br>(0.05)                        |
| Post-Facebook Scandal<br>* Veiled Tradeoff Treatment   | 0.09<br>(0.08)     | -0.01<br>(0.10)            | 0.02<br>(0.10)     | 0.09<br>(0.09)                         |
| Constant   | 0.67***<br>(0.03)  | 0.68***<br>(0.04)          | 0.68***<br>(0.04)  | 0.67***<br>(0.04)                      |
| Observations   | 755                | 689                        | 619                | 734                                    |
| Adjusted $R^2$   | 0.03               | 0.03                       | 0.02               | 0.04                                   |

Table 1.6: Table reports a linear probability regression of a binary variable for whether the participant ended up staying anonymous on binary variables for whether the participant was in the Veiled Tradeoff Treatment, whether the experiment occurred at the height of the Facebook scandal, whether the experiment occurred one month after the Facebook scandal, and interactions between the treatment indicator and date indicators. Each column represents the same regression but with different samples for robustness checks. Block bootstrap standard errors bootstrapped at the treatment level in parentheses. \* $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

In sum, I find no measurable change in survey responses before, during, and after the Facebook scandal, nor do I find any change in behavior in the Direct Tradeoff treatment. I *do*, however, observe that the experimental treatment became significantly less effective, and this was driven by people in the Veiled Tradeoff group opting for privacy more often.

## 5. Discussion

The results of the experiment in this paper provide evidence for two conclusions. First, people do in fact behave inconsistently around privacy decisions. Second, this inconsistency can be explained in part by information avoidance.

Because of unique timing, the paper also sheds some limited light on the effect of public privacy scandals on privacy behavior. The treatment effect dissipated at the height of one of the biggest, most salient privacy scandals of the past decade, but not because people valued privacy more when directly asked. Rather, when the scandal hit, people's ability to take advantage of the costless veil seems to have weakened. But this change did not signal a new normal -- privacy behavior returned to pre-scandal levels within two months of the scandal breaking.

The results also suggest two directions for future scholarship on internet privacy. First, more research is needed to understand *why* people avoid information about privacy. Second, given that information avoidance can explain privacy inconsistency, more thought should be given to existing policy interventions in internet privacy.

An important unresolved question is *why* people avoid information. There are several plausible mechanisms. One is signaling. People care about privacy, but they also care about being the type of person who cares about privacy. This drives a wedge between the direct tradeoff group and the veiled tradeoff group, because members of the veiled tradeoff group can take the monetary bonus without explicitly choosing to give away their data. In this view, people who tap “No” when a browser asks them to share their location may be evincing a sort of phatic preference: their action helps express righteous anger as much as underlying preferences in a small decision where the stakes are low. A second mechanism is that thinking about a probabilistic chance of losing privacy is itself upsetting, as in the model of anxiety in (Koszegi,

2003). A third mechanism is that people do care about privacy, but are also able to turn off their minds to privacy losses that are not directly in front of their face. Another mechanism is choosing costs (Sunstein, 2014). It takes effort to make a decision between money and privacy, especially if privacy costs are inchoate or hard to measure. Perhaps the direct tradeoff group has no choice but to make this effort, but the veiled tradeoff group might rationally decide that it is better to avoid doing the calculations, exploiting the veil as a cognitive shortcut. Still another explanation of the results is that all these mechanisms are true, to greater or lesser degree depending on the person and the context. Future research can explore this, for example by changing whether the decision to hide information is active or passive, or by changing the probabilities in the experiment.

More broadly, the experiment suggests reason for skepticism about policy interventions aimed at improving consumer decision-making with better information. Under the simpler explanations of privacy inconsistency -- revealed preference and ignorance -- policy-makers agree that more and simpler information is better (FTC, 2012; Cranor, 2010). Specifically, better notice means better choices, provided the notice is at low cost. Given this, there have been extensive efforts to improve privacy disclosures, for example with a privacy nutrition label. However, this experiment shows that such efforts will be a steep climb. The results presented here show that even when the privacy settings could be revealed instantly, and even when the settings were a mere two words long ("low privacy" and "high privacy"), most participants still opted not to click. Even when, or especially when, a privacy disclosure is salient and clear and easily accessible, people may have struthious preferences.

## References

- Acquisti, A., Brandimarte, L. & Loewenstein, G. (2015), 'Age of Information,' *Science* 347(6221), 509–515.
- Acquisti, A., John, L. K. & Loewenstein, G. (2013), 'What Is Privacy Worth?', *The Journal of Legal Studies* 42(2), 249–274.
- Andreoni, J. & Sprenger, C. (2012), 'Estimating Time Preferences from Convex Budgets', *American Economic Review* 102(7), 3333–3356.
- Athey, S., Catalini, C. & Tucker, C. (2017), The Digital Privacy Paradox: Small Money, Small Costs, Small Talk.
- Bakos, Y., Marotta-Wurgler, F. & Trossen, D. R. (2014), 'Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts', *The Journal of Legal Studies* 43(1), 1–35.
- Ben-Shahar, O. & Chilton, A. (2016), 'Simplification of Privacy Disclosures: An Experimental Test', *The Journal of Legal Studies* 45(S2), S41–S67.
- Dana, J., Weber, R. A. & Kuang, J. X. (2007), 'Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness', *Economic Theory* 33(1), 67–80.
- DellaVigna, S., List, J. A. & Malmendier, U. (2012), 'Testing for Altruism and Social Pressure in Charitable Giving', *Quarterly Journal of Economics* 127(1), 1–56.
- Exley, C. L. (2016), 'Excusing selfishness in charitable giving: The role of risk', *Review of Economic Studies* 83(2), 587–628.
- Federal Trade Commission, . (2012), Protecting Consumer in an Era of Rapid Change: recommendations for businesses and policymakers, Technical Report March, Federal Trade Commission.
- Golman, R., Hagmann, D. & Loewenstein, G. (2017), 'Information Avoidance', *Journal of Economic Literature* 55(1), 96–135.
- Guttman Institute, . (2018), Requirements for Ultrasound, Technical report.  
URL: <https://www.guttman.org/state-policy/explore/requirements-ultrasound>
- Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C. & Bigham, J. P. (2018), 'A Data-Driven Analysis of Workers ' Earnings on Amazon Mechanical Turk', *Conference on Human Factors in Computing Systems (CHI 2018)* pp. 1–14.
- Hoffman, E., Schwartz, D., Spitzer, M. & Talley, E. (2017), Patently Risky : Framing , Innovation and Entrepreneurial Preferences.
- Irvine, K., Hoffman, D. A. & Wilkinson-Ryan, T. (2018), 'Law and Psychology Grows Up, Goes Online, and Replicates', *Journal of Empirical Legal Studies* 15(2), 320–355.

- Posner, R. A. (1978), 'The right of privacy', *Georgia Law Review* 12(3), 393.
- Prosser, W. L. (1960), 'Privacy', *California Law Review* 48(3).
- Stigler, G. J. (1961), 'The Economics of Information', *Journal of Political Economy* 69(3), 213–225.
- Strahilevitz, L. J. (2010), 'Reunifying Privacy Law', *California Law Review* 98(6), 2007–2048.
- Sullivan, P. S., Lansky, A. & Drake, A. (2004), 'Failure to Return for HIV Test Results Among Persons at High Risk for HIV Infection', *JAIDS Journal of Acquired Immune Deficiency Syndromes* 35(5), 511–518.
- Sunstein, C. R. (2014), 'Choosing Not to Choose', *SSRN Electronic Journal* 5(1999), 1–52.
- Warren, S. D. & Brandeis, L. D. (1890), 'The Right to Privacy', *Harvard Law Review* 4(5), 193–220.



## **Chapter 2: Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment**

## 1. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment<sup>1</sup>

Over the past fifty years, there have been considerable societal efforts to reduce the level of discrimination against African-Americans in the United States. In the context of housing and rental accommodations, antidiscrimination laws have sought to eliminate discrimination through regulation. While racial discrimination continues to exist in rental markets, it has improved in the last two decades (Yinger 1998, U.S. Dep't of Housing and Urban Development, 2012; *compare* Zhao et al., 2005 to Ondrich et al., 1999).

Yet in recent years, markets have changed dramatically, with a growing share of transactions moving online. In the context of housing, Airbnb has created a new market for short-term rentals that did not previously exist, allowing small landlords to increasingly enter the market. Whereas antidiscrimination laws ban the landlord of a large apartment building from discriminating based on race, the prevailing view among legal scholars is that such laws likely do not reach many of the smaller landlords using Airbnb (Belzer & Leong, *forthcoming*; Todisco, 2015).

In this paper, we investigate the existence and extent of racial discrimination on Airbnb, the canonical example of the sharing economy. Airbnb allows hosts to rent out houses, apartments, or rooms within an apartment. To facilitate these transactions, Airbnb promotes properties to prospective guests, facilitates communication, and handles payment and some aspects of customer service. Airbnb allows hosts to decide whether to accept or reject a guest after seeing his or her name and often a picture – a market design choice that may further enable discrimination.

To test for discrimination, we conduct a field experiment in which we inquire about the availability of roughly 6,400 listings on Airbnb across five cities. Specifically, we create guest

---

<sup>1</sup> Co-authored with Michael Luca and Benjamin Edelman

accounts that differ by name but are otherwise identical. Drawing on the methodology of a labor market experiment by Bertrand and Mullainathan (2004), we select two sets of names—one distinctively African-American and the other distinctively White.<sup>2</sup>

We find widespread discrimination against guests with distinctively African-American names. African-American guests received a positive response roughly 42% of the time, compared to roughly 50% for White guests.<sup>3</sup> This 8 percentage point (roughly 16%) penalty for African-American guests is particularly noteworthy when compared to the discrimination-free setting of competing short-term accommodation platforms such as Expedia. The penalty is consistent with the racial gap found in contexts ranging from labor markets to online lending to classified ads to taxicabs.<sup>4</sup>

Combining our experimental results with observational data from Airbnb’s site, we investigate whether different types of hosts discriminate more, and whether discrimination is more common at certain types of properties based on price or local demographics. Our results are remarkably persistent. Both African-American and White hosts discriminate against African-American guests; both male and female hosts discriminate; both male and female African-American guests are discriminated against. Effects persist both for hosts that offer an entire property and for hosts who share the property with guests. Discrimination persists among

---

2 We build on the large literature using audit studies to test for discrimination. Past research considers African-Americans and applicants with prison records in the labor market (Pager 2003), immigrants in the labor market (Oreopoulos 2011), Arabic job-seekers (Carlsson & Rooth 2007), gender (Lahey 2008), long-term unemployment (Ghayad 2014), and going to a for-profit college (Deming et al. 2016), among many others.

3 Some caution is warranted here. We only observe a gap between distinctively white and distinctively African-American names, which differ not only by suggested ethnicity but also potentially by socioeconomic status (Fryer and Levitt, 2004). For ease of exposition, we describe our results in terms of differences among the “African-American guests” or the “white guests,” or use the term “race gap,” without also specifying that our results may better be described as a “race and socioeconomic status gap.” Section 5 discusses this issue in more detail.

4 Doleac & Stein (2013) find a 62% to 56% gap in offer rates for online classified postings. Bertrand and Mullainathan (2004) find a 10% to 6% gap in callback rates for jobs. Pope & Sydnor (2011) find a 9% to 6% gap in lending rates in an online lending market. Ayres et al. (2005) find a 20% to 13% gap in how often taxi drivers receive a tip.

experienced hosts, including those with multiple properties and those with many reviews. Discrimination persists and is of similar magnitude in high and low priced units, in diverse and homogeneous neighborhoods.

Because hosts' profile pages contain reviews (and pictures) from recent guests, we can cross-validate our experimental findings using observational data on whether the host has recently had an African-American guest. We find that discrimination is concentrated among hosts with no African-American guests in their review history. When we restrict our analysis to hosts who have had an African-American guest in the recent past, discrimination disappears – reinforcing the external validity of our main results, and suggesting that discrimination is concentrated among a subset of hosts.

To explore the cost to a host of discriminating, we check whether each listing is ultimately rented for the weekend we inquired about. Combining that information with the price of each listing, we estimate that a host incurs a cost of roughly \$65-\$100 in foregone revenue by rejecting an African-American guest.

Overall, our results suggest a cause for concern. While discrimination has shrunk in more regulated offline markets, it arises and persists in online markets. Government agencies at both the federal and state level have routinely conducted audit studies to test for racial discrimination since 1955 in offline markets. One might imagine implementing regular audits in online markets as well; indeed, online audits might be easier to run at scale due to improved data access and reduced implementation cost.

Our results also reflect the design choices that Airbnb and other online marketplaces use. It is not clear a priori how online markets will affect discrimination. To the extent that online markets can be more anonymous than in-person transactions, there may actually be less room for

discrimination. For example, Ayres and Siegelman (1995) find that African-American car buyers pay a higher price than white car buyers at dealerships, whereas Morton et al. (2003) find no such racial difference in online purchases. Similarly, platforms such as Amazon, eBay, and Expedia offer little scope for discrimination, as sellers effectively pre-commit to accept all buyers regardless of race or ethnicity. However, these advantages are by no means guaranteed, and in fact they depend on design choices made by each online platform. In this situation, Airbnb's design choices enable widespread discrimination.

## **2. ABOUT AIRBNB**

Airbnb is a popular online marketplace for short-term rentals. Founded in 2008, the site gained traction quickly and, as of November 2015, it offers 2,000,000 listings worldwide.<sup>5</sup> This is more than three times as many as Marriott's 535,000 rooms worldwide. Airbnb reports serving over 40 million guests in more than 190 countries.

While the traditional hotel industry is dominated by hotels and inns that each offer many rooms, Airbnb enables anyone to post even a single room that is vacant only occasionally. Hosts provide a wealth of information about each listing, including the type of property (house, apartment, boat, or even castle, of which there are over 1400 listed), the number of bedrooms and bathrooms, the price, and location. Each host also posts information about herself. An interested guest can see a host's profile picture as well as reviews from past guests. Airbnb encourages prospective guests to confirm availability by clicking a listing's "Contact" button to write to the host.<sup>6</sup> In our field experiments (described in the next section), we use that method to evaluate a host's receptiveness to a booking from a given guest.

---

<sup>5</sup> <https://www.airbnb.com/about/about-us>

<sup>6</sup> See "How do I know if a listing is available", <https://www.airbnb.com/help/question/137>.

### **3. Experimental Design**

#### **3.1. Sample and Data Collection**

We collected data on all properties offered on Airbnb in Baltimore, Dallas, Los Angeles, St. Louis, and Washington, D.C. as of July 2015. Our goal was to collect data from the top twenty metropolitan areas from the 2010 census. We started with these five cities because they had varying levels of Airbnb usage and came from diverse geographic regions. Baltimore, Dallas, and St. Louis offer several hundred listings each, while Los Angeles and Washington, D.C. have several thousand. We stopped data collection after these five cities because Airbnb became increasingly rapid in blocking our automated tools which logged into guest accounts and communicated with hosts. (We considered taking steps to conceal our methods from Airbnb, but ultimately declined to do so.)

Because some hosts offer multiple listings, we selected only one listing per host using a random number generator. This helped to reduce the burden on any given host, and it also prevented a single host from receiving multiple identical emails. Each host was contacted for no more than one transaction in our experiment.

We also collected data from each host's profile page. This allowed us to analyze host characteristics in exceptional detail. First, we saved the host's profile image. We then employed Mechanical Turk workers to assess each host image for race (White, African-American, Asian, Hispanic, multiracial, unknown), gender (male, female, two people of the same gender, two people of different genders, unknown), and age (young, middle-aged, old). We hired two Mechanical Turk workers to assess each image, and if the workers disagreed on race or gender, we hired a third to settle the dispute. If all three workers disagreed (as happened, for example, for a host whose profile picture was an image of a sea turtle), we manually coded the picture. We coded race as "unknown"

when the picture did not show a person. Through this procedure, we roughly categorized hosts by race, gender, and age.

Profile pages also revealed other variables of interest. We noted the number of properties each host offers on Airbnb, anticipating that professional hosts with multiple properties might discriminate less often than others. We retrieved the number of reviews the host has received, a rough measure of whether the host is an avid Airbnb user or a casual one. We further checked the guests who had previously reviewed each host. Airbnb posts the photo of each such guest, so we used Face++, a face-detection API, to categorize past guests by race, gender, and age.<sup>7</sup> This allows us to examine relationships between a host's prior experience with African-American guests and the host's rejection of new African-American requests.

We also collected information about each listing. We recorded the price of the listing, the number of bedrooms and bathrooms, the cancellation policy, any cleaning fee, and the listing's ratings from past guests. We also measured whether the listing offered an entire unit versus a room in a larger unit, yielding a proxy for how much the host interacts with the guest.

Each listing included a longitude and latitude, which allowed us to link to census demographic data to assess the relationship between neighborhood demographics and discrimination. After linking the latitude and longitude to a census tract, we used census data on the number of African-American, Hispanic, Asian, and White individuals. Table 2.1 presents summary statistics about the hosts and listings as well as balanced treatment tests.

---

<sup>7</sup> In addition to detecting race, gender, and age, Face++ estimates its confidence for each trait. When Face++ was unable to make a match or its confidence was below 95 out of 100, we used Mechanical Turk, to categorize the past guest via the method described above.

Table 2.1: Summary Statistics

| <i>Variables</i>                      | <i>Mean</i> | <i>Std. Dev.</i> | <i>Obs.</i> | <i>Mean, White Accounts</i> | <i>Mean, Af-Am Accounts</i> | <i>p-value</i> |
|---------------------------------------|-------------|------------------|-------------|-----------------------------|-----------------------------|----------------|
| Host is White                         | 0.63        | 0.48             | 6,392       | 0.64                        | 0.63                        | 0.15           |
| Host is Af-Am                         | 0.08        | 0.27             | 6,392       | 0.08                        | 0.08                        | 0.97           |
| Host is female                        | 0.38        | 0.48             | 6,392       | 0.38                        | 0.37                        | 0.44           |
| Host is male                          | 0.30        | 0.46             | 6,392       | 0.3                         | 0.3                         | 0.90           |
| Price (\$)                            | 181         | 1,280            | 6,302       | 166.43                      | 195.81                      | 0.36           |
| Number of bedrooms                    | 3.18        | 2.26             | 6,242       | 3.18                        | 3.18                        | 0.96           |
| Number of bathrooms                   | 3.17        | 2.26             | 6,285       | 3.17                        | 3.17                        | 0.93           |
| Number of reviews                     | 30.87       | 72.51            | 6,390       | 30.71                       | 31.03                       | 0.86           |
| Host has multiple listings            | 0.16        | 0.36             | 6,392       | 0.32                        | 0.33                        | 0.45           |
| Host has 1+ reviews from Af-Am guests | 0.29        | 0.45             | 6,390       | 0.29                        | 0.28                        | 0.38           |
| Airbnb listings per Census tract      | 9.51        | 9.28             | 6,392       | 9.49                        | 9.54                        | 0.85           |
| % population Af-Am (Census tract)     | 0.14        | 0.2              | 6,378       | 0.14                        | 0.14                        | 0.92           |

We later checked each listing to see whether hosts were ultimately able to fill openings. Our guests inquired about reservations eight weeks in advance. Thus, if a guest sent a message on August 1 about the weekend of September 25, we checked on Friday, September 24 to see whether the specified listing was still listed as available.

### 3.2. Treatment groups



Our analysis used four main treatment groups based on the perceived race and gender of the test guest accounts. Hosts were contacted by guests with names that signaled African-American males, African-American females, White males, and White females, drawn from Bertrand and Mullainathan (2004). The list was based on the frequency of names from birth certificates of babies born between 1974 and 1979 in Massachusetts. Distinctively White names are those that are most likely to be White, conditional on the name, and similarly for distinctively African-American names. To validate the list, we conducted a survey in which we asked participants to quickly categorize each name as White or African-American. With just three seconds permitted for a response, survey takers had little time to think beyond a gut response. The survey results, presented in Appendix Table C.1, confirm that the names continue to signal race.<sup>8</sup>

We then created twenty Airbnb accounts, identical in all respects except for guest names. Our names included ten that are distinctively African-American and ten distinctively White names, divided into five male and five female names within each group. To avoid the confounds that would result from pictures, we use only names; our Airbnb profiles include no picture of the putative guest. From these twenty guest accounts, we sent messages to prospective hosts. Each host was randomly assigned one of our twenty guest accounts. Figure 2.1 presents a representative email from one of our guests to an Airbnb host. The name and dates changed depending on the message sender and when the message was sent.<sup>9</sup> In choosing the dates, we asked hosts about a

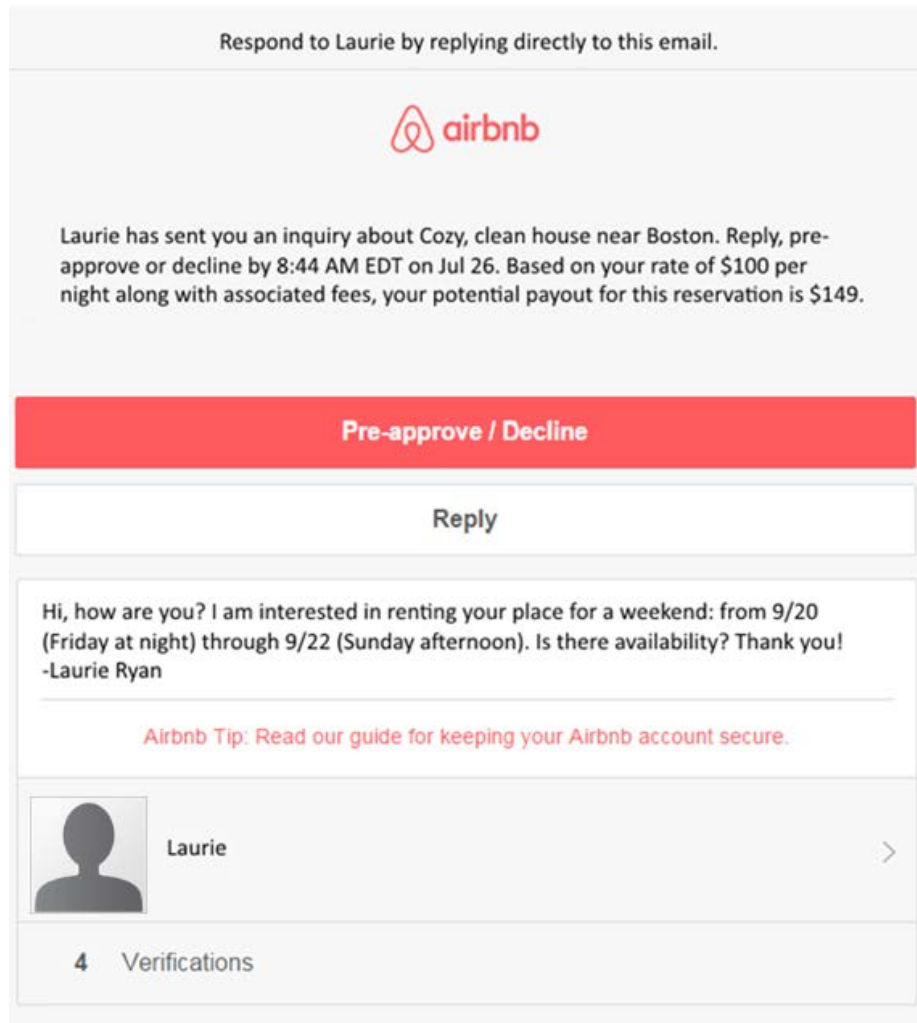
---

<sup>8</sup> On a scale of 0 to 1, where 0 is African-American, the White female names each had an average survey response of 0.90 or above, and the African-American female names all had an average score of 0.10 or below. The male names showed slightly more variation but tell the same story: all the White male names scored 0.88 or above, and all the African-American male names except for Jermaine Jones scored 0.10 or below. The Appendix presents the full results of the survey.

<sup>9</sup> No more than 48 hours elapsed between our first contact to a host in a given city, and the completion of our contacting hosts in that city. Furthermore, no hosts in our sample had listings in more than one of the five cities we tested. Hence, it is unlikely that a host contacted later on in the study would have learned about the experiment.

weekend that was approximately eight weeks distant from when the message was sent. We limited our search to those properties that were listed as available during the weekend in question.

Figure 2.1: Sample Treatment



### 3.3. Experimental Procedure

We sent roughly 6,400 messages to hosts between July 7, 2015 and July 30, 2015.<sup>10</sup> Each message inquired about availability during a specific weekend in September. When a host replied

<sup>10</sup> Our initial goal was to collect roughly 10,000 responses. This was based on a power analysis, which in turn used an effect size calculated from Edelman and Luca (2014). To find a similar effect size, we would need a sample size of roughly 3,000 hosts. But, to calculate an effect among a subgroup of hosts, like African-American hosts, which

to a guest, we replied to the host with a personal message clarifying that we (as the guest) were still not sure if we would visit the city or if we would need a place to stay. We sent this reply in order to reduce the likelihood of a host holding inventory for one of our hypothetical guests.

We tracked host responses over the 30 days that followed each request. A research assistant then coded each response into categories. The majority of responses were in one of six groups: “No response” (if the host did not respond within 30 days); “No or listing is unavailable”; “Yes”; “Request for more information” (if the host responded with questions for the guest); “Yes, with questions” (if the host approved the stay but also asked questions); “Check back later for definitive answer”; and “I will get back to you.” As these categories show, our initial categorizations used subtle distinctions between possible responses. In our analyses below, however, we restrict our attention to the simplest response – “Yes” – though all of our results are robust to using “No” instead, as well as to ignoring non-responses or to using broader definitions of “Yes.”

We collected all data using scrapers we built for this purpose. We sent inquiries to Airbnb hosts using web browser automation tools we built for this purpose. Our Institutional Review Board approved our methods before we began collecting data.

#### **4. Results**

Table 2.2 presents the main effect. We find that inquiries from guests with White-sounding names are accepted roughly 50% of the time. In contrast, guests with African-American-sounding names are accepted roughly 42% of the time. Columns 2 and 3 introduce additional control variables related to the host or the property. The effect stays constant at a roughly eight percentage point gap across these specifications, controlling for the host’s gender, race, an indicator for

---

represent roughly 7% of the Airbnb population, we would need a sample size closer to 10,000. We fell short of this goal for an exogenous reason: Airbnb shut down the experimental accounts after we collected roughly 6,400 responses.

whether the host has multiple listings, an indicator for whether the property is shared, host experience (whether the host has more than ten reviews), and the log of the listing price.

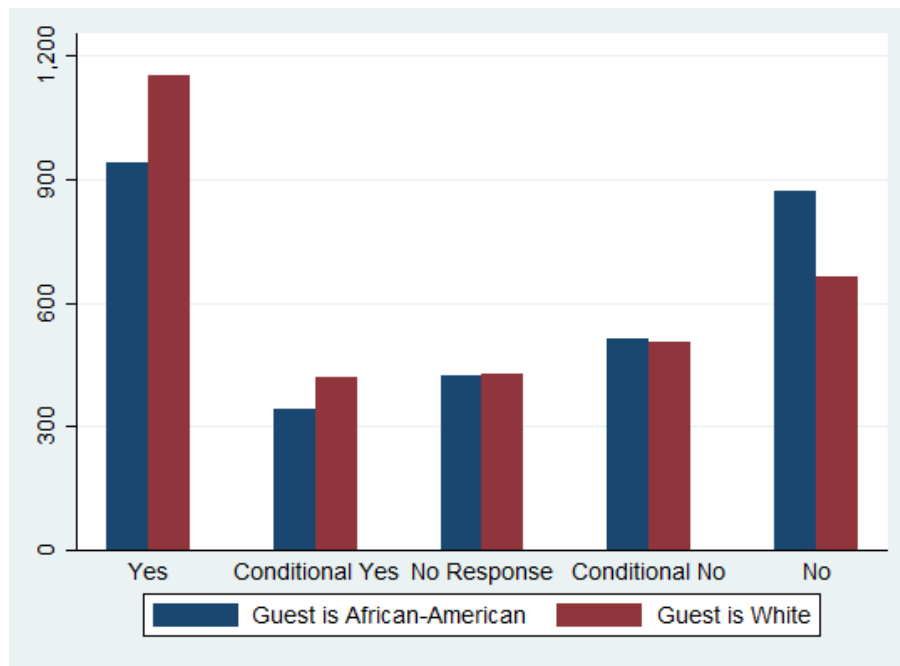
Table 2.2. The Impact of Race on Likelihood of Acceptance

|                            | <i>Dependent Variable: 1(Host Accepts)</i> |                                |                                |
|----------------------------|--|--------------------------------|--------------------------------|
| Guest is African-American  | -0.08 <sup>***</sup><br>(0.02)             | -0.08 <sup>***</sup><br>(0.02) | -0.09 <sup>***</sup><br>(0.02) |
| Host is African-American   |  | 0.07 <sup>**</sup><br>(0.02)   | 0.09 <sup>***</sup><br>(0.02)  |
| Host is Male               |  | -0.05 <sup>***</sup><br>(0.01) | -0.05 <sup>***</sup><br>(0.01) |
| Host has Multiple Listings |  |                                | 0.09 <sup>***</sup><br>(0.02)  |
| Shared Property            |  |                                | -0.07 <sup>***</sup><br>(0.02) |
| Host has 10+ Reviews       |  |                                | 0.12 <sup>***</sup><br>(0.01)  |
| ln(Price)                  |  |                                | -0.06 <sup>***</sup><br>(0.01) |
| Constant                   | 0.49 <sup>***</sup><br>(0.01)              | 0.50 <sup>***</sup><br>(0.01)  | 0.76 <sup>***</sup><br>(0.07)  |
| Observations               | 6,235                                      | 6,235                          | 6,168                          |
| Adjusted $R^2$             | 0.006                                      | 0.009                          | 0.040                          |

Notes: A host's response is coded as a "Yes" if, in her reply to the guest, she invites the guest to stay at the property, if she offers a special deal ("book within 24 hours and get a discount"), or approves the guest while also asking some clarifying question ("You can stay, but how many people will you have with you?"). Standard errors are clustered by (guest name)\*(city) and are reported in parentheses. \* p < .10. \*\* p < .05. \*\*\* p < .01.

As noted, we break down hosts' responses into 11 categories. Figure 2.2 shows the frequency of each response by race. One might worry that results are driven by differences in host responses that are hard to classify, such as conditional "Yes" responses. Similarly, we would be concerned if our findings were driven by differences in response rate. African-American accounts might be more likely to be categorized as spam, or hosts may believe that African-American accounts are more likely to be fake, in which case one might expect higher non-response rates for African-American accounts. But as Figure 2.2 shows, the discrimination results occur because of differences in simple "Yes" or "No" responses, not because of non-responses or intermediate responses (like a conditional "Yes").

Figure 2.2: Host Responses by Race



In the rest of this section, we use the wealth of data available on Airbnb about the host and location for each listing to look for factors that influence the gap between white

and African-American names. Does the identity of the host matter? Does the location of the property matter? Generally, we find that the discrimination is remarkably robust.

#### **4.1. Effects by Host Characteristics**

We first check whether our finding changes based on the identity of the host. If discrimination is driven by homophily (in-group bias), then the host's race should matter. According to this theory, hosts might simply prefer guests of the same race. If homophily were the primary factor driving differential guest acceptance rates, then African-American guests would face higher acceptance rates from African-American hosts. Table 2.3 presents regressions that include guest race, host race, and an interaction term. Across the entire sample of hosts, the interaction between the race and guest of the host is not significantly different from zero, but the point estimate is noisy. This result masks heterogeneity across genders. Columns 2 and 3 of Table 2.3 report the same regression limited to male hosts and female hosts, respectively. Among male hosts, the interaction between the host's race and guest's race shows a widening of the race gap by 11 percentage points, whereas among females, the race gap narrows by 11 percentage points. Both estimates are noisy; we cannot reject coefficients of zero.<sup>11</sup>

---

<sup>11</sup> Table 2.4 explores the effect of the host's race with more nuance. It shows the proportion of Yes responses from each gender/race cell among hosts in response to each gender/race cell among guests. African-American male hosts discriminate against African-American male and female guests. White hosts of both genders are more likely to accept white guests of either gender. African-American female hosts are the only exception: they accept African-American female guests more than any other group. Thus, with the exception of African-American females, the data is inconsistent with homophily.

Table 2.3: Race Gap by Race of the Host

|  | <i>Dependent Variable: 1(Host Accepts)</i> |                    |                    |                   |
|--|--|--------------------|--------------------|-------------------|
|  | All  | Male               | Female             | Other             |
|  | Hosts                                      | Hosts              | Hosts              | Hosts             |
| Guest is African-American                                  | -0.08***<br>(0.02)                         | -0.09***<br>(0.02) | -0.09***<br>(0.02) | -0.07*<br>(0.03)  |
| Host is African-American                                   | 0.06*<br>(0.03)                            | 0.19***<br>(0.05)  | -0.00<br>(0.04)    | 0.03<br>(0.09)    |
| Host is African-American *<br>Guest is African-American    | 0.01<br>(0.05)                             | -0.11<br>(0.08)    | 0.11<br>(0.06)     | -0.06<br>(0.14)   |
| Constant   | 0.48***<br>(0.01)                          | 0.44***<br>(0.02)  | 0.50***<br>(0.02)  | 0.50***<br>(0.02) |
| Observations   | 6235                                       | 1854               | 2336               | 2045              |
| Adjusted $R^2$   | 0.007                                      | 0.015              | 0.007              | 0.003             |
| Implied Coefficient on Guest<br>is Af-Am + Host is Af-Am * | -0.07<br>(0.05)                            | -0.19**<br>(.08)   | 0.02<br>(.06)      | -0.12<br>(0.14)   |

Notes: Other hosts are hosts we could not classify as male or female. Of the 2,045 host pictures we could not classify for gender, 972 had a picture of a mixed-gender couple, 259 had a same-gender couple, 603 had a picture without a human in it, and the rest could not be classified. Standard errors are clustered by (guest name)\*(city) and are reported in parentheses. \*  $p < .10$ . \*\*  $p < .05$ . \*\*\*  $p < .01$ .

Discrimination may also be influenced by a host's proximity to the guest. For example, Becker (1957) formalizes racial discrimination as distaste for interactions with individuals of a certain race. On Airbnb, a host must classify each listing as offering an entire unit, a room within a unit, or a shared room. We classify anything other than an entire unit as a "shared property." Column 1 of Table 2.5 shows that the race gap is

roughly the same whether or not a property is shared. (In unreported results, we find that the race gap stays roughly the same in shared properties with only one bathroom.)

Table 2.4. Are Effects Driven by Host Characteristics?

| <i>Dependent Variable: 1(Host Accepts)</i>                          |                        |                        |                        |                        |                    |
|---|------------------------|------------------------|------------------------|------------------------|--------------------|
| Guest is African-American   | -0.07***<br>(0.02)     | -0.08***<br>(0.02)     | -0.09***<br>(0.02)     | -0.11***<br>(0.02)     | -0.09***<br>(0.02) |
| Shared Property   | 0.00<br>(0.01)         |                        |                        |                        |                    |
| Shared Property * Guest is African-American                         | -0.02<br>(0.03)        |                        |                        |                        |                    |
| Host has Multiple Listings  |                        | 0.14***<br>(0.02)      |                        |                        |                    |
| Host has Multiple Listings * Guest is Af-Am                         |                        | -0.01<br>(0.03)        |                        |                        |                    |
| Host has 10+ Reviews  |                        |                        | 0.14***<br>(0.02)      |                        |                    |
| Host has Ten+ Reviews * Guest is Af-Am                              |                        |                        | 0.01<br>(0.02)         |                        |                    |
| Host has 1+ reviews from an Af-Am guest                             |                        |                        |                        |                        | 0.10***<br>(0.01)  |
| Host has 1+ reviews from an Af-Am guest * Guest is Af-Am            |                        |                        |                        |                        | 0.06*<br>(0.02)    |
| Constant  | 0.49***<br>(0.01)      | 0.46***<br>(0.01)      | 0.42***<br>(0.01)      | 0.50***<br>(0.01)      | 0.46***<br>(0.01)  |
| Observations  | 6,235                  | 6,235                  | 6,235                  | 6,235                  | 6,235              |
| Adjusted R <sup>2</sup>   | 0.006                  | 0.014                  | 0.027                  | 0.011                  | 0.019              |
| Implied Coefficient on Guest is Af-Am + Host Trait * Guest is Af-Am | -<br>0.09***<br>(0.02) | -<br>0.09***<br>(0.03) | -<br>0.08***<br>(0.02) | -<br>0.08***<br>(0.03) | -0.04<br>(0.03)    |

Notes: This table presents a linear regression of [Host Accepted] on a Host Trait, the Guest's Race, and an interaction of the two. Standard errors are clustered by (guest name)\*(city) and are reported in parentheses.

\* p < .10. \*\* p < .05. \*\*\* p < .01



One might expect a distinction between casual Airbnb hosts who occasionally rent out their homes, versus professional hosts who offer multiple properties. Roughly a sixth of Airbnb hosts manage multiple properties, and roughly 40% of hosts have at least 10 reviews from past guests. Columns 2 and 3 explore the extent of discrimination among hosts with multiple locations, and those with more than 10 reviews. Across these specifications, the race gap persists with roughly the same magnitude.

Table 2.5. Proportion of Positive Responses by Race and Gender

|                           |                                | <i>Guest Race / Gender</i> |                              |                 |                                |
|---------------------------|--------------------------------|----------------------------|------------------------------|-----------------|--------------------------------|
|                           |                                | White<br>Male              | African-<br>American<br>Male | White<br>Female | African-<br>American<br>Female |
| <i>Host Race / Gender</i> | White<br>Male                  | 0.42                       | 0.35                         | 0.49            | 0.32***                        |
|                           | African-<br>American<br>Male   | 0.64**                     | 0.40                         | 0.59            | 0.43                           |
|                           | White<br>Female                | 0.46                       | 0.35                         | 0.49            | 0.44                           |
|                           | African-<br>American<br>Female | 0.43                       | 0.38                         | 0.53            | 0.59***                        |

Notes: This table shows the proportion of Yes responses by hosts of a certain race/gender to guests of a certain race/gender. \*  $p < .10$ . \*\*  $p < .05$ . \*\*\*  $p < .01$ . P-values from testing that proportion of Yes responses in a specific cell is equal to the proportion of Yes responses from the other cells in that row.

## 4.2. Effects by Listing Characteristics

Just as discrimination was robust across host characteristics, we find that discrimination does not vary based on the cost or location of the property. Column 1 of Table 2.6 shows that, overall, listings above the median price are more likely to reject inquiries. However, discrimination remains both among more expensive and less expensive listings.

Table 2.6. Are Effects Driven by Location Characteristics?

|  | <i>Dependent Variable=1(Host Accepts)</i> |                    |                     |                   |
|--|---|--------------------|---------------------|-------------------|
| Guest is Af-Am   | -0.09***<br>(0.02)                        | -0.08***<br>(0.02) | -0.09***<br>(0.02)  | -0.12**<br>(0.06) |
| Price > Median   | -0.07***<br>(0.02)                        |                    |                     |                   |
| Guest is Af-Am *<br>(Price > Median)                                 | 0.01<br>(0.03)                            |                    |                     |                   |
| Share of Af-Am Population<br>in Census Tract                         |   | 0.05<br>(0.05)     |                     |                   |
| Guest is Af-Am * (Share of<br>Af-Am Population in Census<br>Tract)   |   | 0.02<br>(0.08)     |                     |                   |
| Airbnb Listings per Census<br>Tract                                  |   |                    | -0.0007<br>(0.0009) |                   |
| Guest is Af-Am *<br>(Airbnb Listings per Census<br>Tract)            |   |                    | 0.0008<br>(0.001)   |                   |
| Probability Listing is Filled 8<br>Weeks Later                       |   |                    |                     | 0.56***<br>(0.08) |
| Guest is Af-Am *<br>(Probability Listing is Filled<br>8 Weeks Later) |   |                    |                     | 0.09<br>(0.12)    |
| Constant   | 0.52***<br>(0.02)                         | 0.48***<br>(0.01)  | 0.49***<br>(0.02)   | 0.24***<br>(0.03) |
| Observations   | 6235                                      | 6223               | 6235                | 6101              |
| Adjusted $R^2$   | 0.01                                      | 0.006              | 0.006               | 0.030             |

Notes: Standard errors are clustered by (guest name)\*(city) and are reported in parentheses.

\* p < .10. \*\* p < .05. \*\*\* p < .01.

We can also check whether the listing was eventually filled (for the nights in question) to create a proxy for the desirability of the listing. First, we fit a Probit model to predict the likelihood that the listing was filled, controlling for a fixed city effect and a host of covariates.<sup>12</sup> Then we assign each listing a probability of being filled. This lets us test whether discrimination changes based on the listing's desirability.<sup>13</sup> It does not.

We also hypothesized that the extent of discrimination might vary with the diversity of a neighborhood. More generally, one might expect that geography matters and that discrimination is worse in some areas than others, due to market structure or underlying rates of discrimination among a population. Merging data on neighborhoods by census tract, Column 2 shows that the extent of discrimination does not vary with the proportion of nearby residents who are African-American. Column 3 shows that discrimination is ubiquitous: it does not vary with the number of Airbnb listings within the census tract. We also find discrimination in all cities in our sample, as shown in Appendix Table C.2.

#### **4.3. Robustness – Effects by Name**

Table 2.7 shows the proportion of positive responses broken down by name. The effect is robust across choice of names. For example, the African-American female name with the most positive responses (Tamika) received fewer positive responses than the White female name with the fewest positive responses (Kristen), though this difference is not statistically significant. Similarly, the African-American males with the most positive

---

<sup>12</sup> The covariates are as follows: the host's race and gender, the price, number of bedrooms, whether the property is shared, whether the bathroom is shared, the number of reviews, the age of the host, whether the host operates multiple listings, the proportion of White people in the census tract, and the number of Airbnb listings in the census tract.

<sup>13</sup> We thank an anonymous reviewer for suggesting this approach.

responses (Darnell and Rasheed) received fewer acceptances than the White male with the fewest positive responses (Brad).

Table 2.7. Proportion of Positive Responses, by Name

| Entire Sample       |                 | 0.43<br>(6,390)                |                 |
|---------------------|-----------------|--------------------------------|-----------------|
| <i>White Female</i> |                 | <i>African-American Female</i> |                 |
| Allison Sullivan    | 0.49<br>(306)   | Lakisha Jones                  | 0.42<br>(324)   |
| Anne Murphy         | 0.56**<br>(344) | Latonya Robinson               | 0.35**<br>(331) |
| Kristen Sullivan    | 0.48<br>(325)   | Latoya Williams                | 0.43<br>(327)   |
| Laurie Ryan         | 0.50<br>(327)   | Tamika Williams                | 0.47**<br>(339) |
| Meredith O'Brien    | 0.49<br>(303)   | Tanisha Jackson                | 0.40<br>(309)   |
| <i>White Male</i>   |                 | <i>African-American Male</i>   |                 |
| Brad Walsh          | 0.41*<br>(317)  | Darnell Jackson                | 0.38<br>(285)   |
| Brent Baker         | 0.48<br>(332)   | Jamal Jones                    | 0.33<br>(328)   |
| Brett Walsh         | 0.44<br>(279)   | Jermaine Jones                 | 0.36<br>(300)   |
| Greg O'Brien        | 0.45<br>(312)   | Rasheed Jackson                | 0.38<br>(313)   |
| Todd McCarthy       | 0.43<br>(314)   | Tyrone Robinson                | 0.36<br>(254)   |

Notes: The table reports the proportion of Yes responses by name. The number of messages sent by each guest name is shown in parentheses. \*  $p < .10$ . \*\*  $p < .05$ . \*\*\*  $p < .01$ . P-values from test of proportion. Null hypothesis is that the proportion of Yes responses for a specific name are equal to the proportion of Yes responses for all other names of the same race\*gender cell.

#### 4.4. Comparing Experimental Results with Observational Patterns

Each listing page includes reviews from previous guests, along with profile pictures for these guests. This allows us to see which hosts previously accepted African-American guests (although not all guests leave reviews and not all guests have photos that reveal their race). We use this data to assess the external validity of our results.

We collected profile pictures from the ten most recent reviews on each listing page. We categorized these past guests by race and gender, finding that 29% of hosts in our sample had at least one review from an African-American guest. We then regressed the likelihood of a host responding positively to our inquiry on the race of the guest, whether the host has at least one recent review from an African-American guest, and an interaction between these variables. Column 5 of Table 2.5 reports the results. We find that the race gap drops sharply among hosts with at least one recent review from an African-American guest. We cannot reject zero difference for requests from our African-American test accounts versus requests from our White test accounts, though this result is only significant at the 10% level.<sup>14</sup>

This finding reinforces our interpretation of our main effects, including the role of race and the interpretation that observed differences reflect racial discrimination by Airbnb hosts. Put another way, if our findings are driven by a quirk of our experimental design,

---

<sup>14</sup> These findings are robust to alternative specifications of a host's past guests. The same substantive results hold if we look at the raw number of reviews from African-Americans, rather than whether there is at least one such review. The same is true if we use the proportion of reviews from African-American guests.

rather than race, then it is difficult to explain why the race gap disappears precisely among hosts with a history of accepting African-American guests.

#### **4.5. Importance of Profile Pictures and More Complete Profiles**

A related concern is that we used guest profiles that were relatively bare. A host may hesitate to accept a guest without a profile picture or past reviews. Of course, this alone cannot explain the race gap, since both white and African-American guests had bare profiles. But it does raise the question of whether more complete profiles could mitigate discrimination.<sup>15</sup>

Internal data from Airbnb and observational data on Airbnb users both suggest that profile pictures alone are unlikely to make much difference. With access to internal Airbnb data, Fradkin (2015) looks at roughly 17,000 requests sent to hosts and finds that guests are rejected 49% of the time. Notably, these requests from ordinary Airbnb users, with typical Airbnb profiles, were rejected at a rate similar to that of our guests. In our experiment, as detailed in Appendix Table C.3, 44% of guests were rejected or received no response. Another 11% received a message from a host requesting more information. The remaining 46% were accepted. The similarity in rejection rates suggests that incompleteness of our guests' profiles is not likely to be causing a change in the rejection rate, and reinforces the ecological validity of our experimental design.

---

<sup>15</sup> Similarly, our experiment does not assess whether discrimination occurs because of race or social class. Hanson & Hawley (2011) find, in a field experiment on Craigslist's housing market using similar methodology, that renters with African-American names face a penalty, but that the penalty decreases if the email sent to a landlord signals higher social class. Under some specifications, African-Americans face a statistically significant penalty based on race and an additional penalty for signaling low class, also statistically significant. Under other specifications, the racial gap is not statistically significant when comparing white and African-American guests who both signal high social class. On the whole, the paper indicates that social class and race both play a role.

Other methods indicate that profile pictures seem to have little impact on acceptance decisions. In a logistic regression estimating the probability of receiving a rejection from a host, again using internal Airbnb data, Fradkin (2015) finds that including a profile picture has no significant effect. This matches the observational data we collect: in a random selection of Airbnb users, we found that only 44% have a profile picture. The proportion of guests with a profile picture is higher among users who have left a review, but nonetheless both analyses indicate that the existence of profile pictures plays a small role in host decision-making. Further, even if profile pictures impact rejection rates, it is not clear that the impact should be differential by race. For example, one might expect that pictures would make a guest's race more salient. If our results are driven by race, then our findings would be a lower bound on the true effect.

One limitation of our experiment is that we do not observe the effect of past reviews on discrimination. If our findings are driven by statistical discrimination, positive reviews from previous hosts may reduce the extent of discrimination. However, three factors suggest that reviews are an incomplete response to a discrimination problem. First, our acceptance rates are similar to overall acceptance rates on Airbnb (Fradkin 2015), which indicates that hosts are not treating our test guest accounts differently for lack of reviews, meaning that reviews would be unlikely to eliminate discrimination. Indeed, for reviews to eliminate discrimination, they would need to provide a 16 percent differential increase in acceptance rates for African-Americans, relative to White guests. Second, all Airbnb users necessarily start without past reviews, so a review system would not address any initial barriers to entry that guests face. Third, a subjective review system can itself allow or facilitate discrimination. (*See, e.g.,* Goldin and Rouse, 2000, finding that visually

confirming a musician's gender may influence an expert's judgment of her work.) Whatever mechanism is causing a lower acceptance rate for the African-American guests may also cause a worse rating.

#### 4.6. How much does discrimination cost hosts?

A host incurs a cost for discriminating when rejecting a guest causes a unit to remain empty. The expected cost depends on the likelihood of the property remaining vacant, which in turn depends on the thickness of the market. If a host can easily find a replacement guest, then discrimination is nearly costless for the host. But if a property remains vacant after the host rejects a guest, then discrimination imposes a more significant cost. In other words, the impact on net revenue from discriminating depends on the likelihood of filling a unit with someone of the host's preferred race after rejecting a guest of a disfavored race.

Because we collect data about each property's availability after a host declines a guest, we can estimate the cost in net revenue from discrimination. Suppose a host charges price  $p$  for a listing and pays listing fees  $f$  to Airbnb. Let  $\pi_{replace}$  be the probability of filling the property after rejecting a guest in our study. Then the cost in net revenue of discrimination is as follows:

$$\Delta \text{Net Revenue} = (p - f) - \pi_{replace} \cdot (p - f) = (1 - \pi_{replace}) \cdot (p - f)$$

That is, the cost of discrimination, in terms of net revenue, is the revenue that the host forgoes if the listing remains empty multiplied by the probability that the listing remains empty.



In our data, hosts who rejected or never responded to our inquiries had properties with a median price of \$163 and a mean price of \$295.<sup>16</sup> The numbers are similar and slightly higher if we restrict the sample further to those hosts who rejected African-American guests, or if we expand the sample to hosts who responded positively Yes to our accounts.<sup>17</sup> Airbnb charges each host a fee equal to 3% of the listing price.

After our inquiries, roughly 25.9% of the listings in our study remained vacant on the dates we requested after rejecting or not responding to one of our guests. Another 37.9% remained listed but were no longer available on those dates, suggesting that the host either found another guest or decided to no longer make the property available on the specified dates. The remaining 36.1% of properties were no longer listed on Airbnb. Because it is unclear whether the hosts who exit should be excluded from the sample or treated as not having found a replacement, we develop two estimates.

If we exclude these disappearing hosts from our calculation, 59.4% of hosts found a replacement guest. Setting  $p$  equal to the median price (\$163) and fees at 3% of the median price:

$$\Delta \text{Net Revenue} = (1 - .594) \cdot (\$163 - .03 \cdot \$163) \approx \$64.19$$

If we treat disappearing listings as vacancies, in effect assuming that the host of a dropped listing was not able to find a replacement guest, then only 37.9% of hosts found a replacement guest. The cost of discrimination rises as a result:

---

<sup>16</sup> In calculating price, we sum the listing price and any cleaning fee.

<sup>17</sup> An anonymous reviewer correctly points out that the host we are interested in is the host on the margin of discriminating. But there are hosts far from this margin both within the group of hosts who said Yes and within the group of hosts who said No. Nonetheless, our calculations in this section are not sensitive to which group of hosts we include. When including hosts who said Yes, the median price drops from \$163 to \$150, and the probability of finding a replacement guest rises to 64% instead of 59.4% (excluding disappearing hosts) or 45% instead of 37.9% (including disappearing hosts). Thus, the cost of discrimination drops by about \$10 or \$20 among hosts who say Yes, and therefore either did not discriminate against the African-American accounts or did not get a chance to do so.

$$\Delta \text{Net Revenue} = (1 - .379) \cdot (\$163 - .03 \cdot \$163) \approx \$98.19$$

In this analysis, we focus on the net revenue, which does not incorporate the marginal cost of each night the listing is rented, since we do not directly observe costs. The cost of hosting includes various types of host effort or wear-and-tear to the property. In principle, hosting also entails a risk of damage by a guest, though throughout the relevant period Airbnb automatically provided all hosts with property insurance, which reduces the risk. Our calculation also excludes unobserved benefits of hosting, such as the possibility that a positive review draws more guests in the future and improves the listing position on Airbnb. A full estimate of profit would also need to consider the time cost of looking for new guests after rejecting someone on the basis of race.<sup>18</sup>

While these estimates are clearly noisy, they suggest that hosts incur a real cost by discriminating. The median host who rejects a guest because of race is turning down between \$65 and \$100 of revenue.

## 5. Discussion

Online platforms such as Airbnb create new markets by eliminating search frictions, building trust, and facilitating transactions (Lewis 2011, Luca forthcoming). With the rise of the sharing economy, however, comes a level of discrimination that is impossible in the online hotel reservations process. Clearly, the manager of a Holiday Inn cannot

---

<sup>18</sup> Our calculation also ignores other factors that cut in both directions. Responding with a Yes to a guest does not provide 100% certainty of a paid booking; the guest may choose another option or may not make the trip. In that case, our estimates overstate the revenue loss. Similarly, we have imperfect information about whether a host found a replacement guest. Among other complexities, our guests requested two-night stays; we treat a host as having filled a listing if the host found a replacement guest for at least one of the nights, though a host who filled only one of the nights has nonetheless lost one night of revenue.

examine names of potential guests and reject them based on race or socioeconomic status or some combination of the two. Yet, this is commonplace on Airbnb, which now accounts for a growing share of the short-term rental market.

Our results contribute to a small but growing body of literature suggesting that discrimination persists—and we argue may even be exacerbated—in online platforms. Edelman and Luca (2014) show that African-American hosts on Airbnb seek and receive lower prices than White hosts, controlling for the observable attributes of each listing. Pope and Sydnor (2011) find that loan listings with pictures of African-Americans on Prosper.com are less likely to be funded than similar listings with pictures of White borrowers. Doleac and Stein (2013) show that buyers are less likely to respond to Craigslist listings showing an iPod held by a Black hand compared to an identical ad with a White hand. In contrast, Morton et al. (2003) find no difference by race in price paid for cars in online purchases—a sharp contrast to traditional channels (*see, e.g.*, List, (2004); Zhao et al., (2005)).

One important limitation of our experiment is that we cannot identify the mechanism causing worse outcomes for guests with distinctively African-American names. Prior research shows that distinctively African-American names are correlated with lower socioeconomic status (Fryer and Levitt, 2004). Our findings cannot identify whether the discrimination is based on race, socioeconomic status, or a combination of these two. That said, we note that discrimination disappears among hosts who have previously accepted African-American guests. One might worry that discrimination against our test guest accounts results from our choice of names and hence does not represent patterns that affect genuine Airbnb guests. However, we find that discrimination is limited to hosts who

have never had an African-American guest, which suggests that our results are consistent with any broader underlying patterns of discrimination.

Similarly, our experiment does not provide a sharp test of alternative models of discrimination. The theoretical literature on discrimination often distinguishes between statistical and taste-based discrimination. While our experimental design cannot reject either mechanism, our findings suggest a more nuanced story than either of the classic models. For one, we find homophily among African-American females, but not among other race/gender combinations. Furthermore, we find that discrimination is not sensitive to a measure of proximity between the host and guest. Both findings are in tension with pure taste-based discrimination. But we also find some evidence against pure statistical discrimination. As noted above, we find that hosts who have had an African-American guest in the past exhibit less discrimination than other hosts. This suggests that, at the very least, hosts are using different statistical models as they evaluate potential guests.

### **5.1. Designing a Discrimination-free Marketplace**

Because online platforms choose which information is available to parties during a transaction, they can prevent the transmission of information that is irrelevant or potentially pernicious. Our results highlight a platform's role in preventing discrimination or facilitating discrimination, as the case may be. If a platform aspires to provide a discrimination-free environment, its rules must be designed accordingly.

Airbnb has several options to reduce discrimination. For example, it could conceal guest names, just as it already prevents transmission of email addresses and phone numbers so that guests and hosts cannot circumvent Airbnb's platform and its fees. Communications

on eBay's platform have long used pseudonyms and automatic salutations, so Airbnb could easily implement that approach.

Alternatively, Airbnb might further expand its "Instant Book" option, in which hosts accept guests without screening them first. Closer to traditional hotels and bed and breakfasts, this system would eliminate the opportunity for discrimination. This change also offers convenience benefits for guests, who can count on their booking being confirmed more quickly and with fewer steps. However, in our sample, only a small subset of hosts currently allow instant booking. Airbnb could push to expand the use of this feature, which would also serve the company's broader goal of reducing search frictions.

More generally, our results suggest an important tradeoff for market designers, who set the rules of online platforms, including the pricing mechanisms (Einav et al 2013) and the information that is available and actionable at the time of transaction (Luca forthcoming). Market design principles have generally focused on increasing the information flow within a platform (Bolton et al 2013, Che and Horner 2014, Dai et al 2014, Fradkin et al 2014), but we highlight a situation in which platforms may be providing too much information.

## **5.2. Policy Implications**

Because the legal system grants considerable protection to online marketplaces, Airbnb is unlikely to be held liable for allowing discrimination on its platform. Within the United States, the Civil Rights Act of 1964 prohibits discrimination in hotels (and other public accommodations) based on race, color, religion, or national origin. But these laws appear to be a poor fit for the informal sharing economy, where private citizens rent out a room in their home (Belzer and Leong, *forthcoming*; Todisco, 2015). As discussed in

Edelman and Luca (2014), any changes by Airbnb would likely be driven by ethical considerations or public pressure rather than law. In contrast, offline rental markets and hotels have been subject to significant regulation (as well as audit studies to test for discrimination) for decades. This contributes to worry among policy-makers that online short-term rental markets like Airbnb may be displacing offline markets, which are more heavily regulated (Schatz et al, 2016). One clear policy implication is that regulators may want to audit Airbnb hosts using an approach based on our paper—much like longstanding efforts to reduce discrimination in offline rental markets.

One might have hoped that online markets would cure discrimination, and it seems a different design might indeed do so. Regrettably, our analysis indicates that at Airbnb, this is not yet the case.

## References

- Ayres, I., & Siegelman, P. (1995). Race and Gender Discrimination in Bargaining for a New Car. *American Economic Review*, 85(3), 304–321.
- Ayres, I., F. Vars, & N. Zakariya. (2005). To Insure Prejudice: Racial Disparities in Taxicab Tipping. *Yale Law Journal*, 114(7), 1613-1674.
- Becker, G. (1957). The Economics of Discrimination. *The University of Chicago Press*.
- Belzer, A., & Leong, N. (2017). The New Public Accommodations. *Georgetown Law Journal*, 105 (forthcoming).
- Benner, K. (2016). Airbnb Adopts Rules to Fight Discrimination by Its Hosts. *New York Times*, A1, available at <http://www.nytimes.com/2016/09/09/technology/airbnb-anti-discrimination-rules.html>.
- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4), 991–1013.
- Bolton, G., B. Greiner, and A. Ockenfels. (2013). Engineering Trust: Reciprocity in the Production of Reputation Information. *Management Science*, 59(2), 265-285.
- Carlsson, M., & Rooth, D. (2007). Evidence of Ethnic Discrimination in the Swedish Labor Market Using Experimental Data. *Labour Economics*, 14(4), 716-729.
- Che, Y. and J. Hörner. (2014). Optimal Design for Social Learning. *Working Paper*.
- Dai, W., G. Jin, J. Lee, & M. Luca. (2014). Optimal Aggregation of Consumer Ratings: an Application to Yelp.com. *NBER Working Paper*.

Deming, D. J., Yuchtman, N., Abulafi, A., Golding, C., & Katz, L. F. (2016). The Value of Postsecondary Credentials in the Labour Market: An Experimental Study. *American Economic Review*, 106(3), 778-806.

Doleac, J., & L. Stein. (2013). The Visible Hand: Race and Online Market Outcomes. *Economic Journal*, 123(572), 469-492.

Edelman, B, and M. Luca. (2014). Digital Discrimination: The Case of Airbnb.com. *Harvard Business School Working Paper*.

Einav, L., C. Farronato, J. Levin, & N. Sundaresan. (2013). Sales Mechanisms in Online Markets: What Happened to Internet Auctions? *Working paper*.

Finley, T. (2016). These Airbnb Alternatives Want To Make Travel More Welcoming For Black People. *The Huffington Post*. Available at [http://www.huffingtonpost.com/entry/inclusive-airbnb-alternatives\\_us\\_5768462ae4b0853f8bf1c675](http://www.huffingtonpost.com/entry/inclusive-airbnb-alternatives_us_5768462ae4b0853f8bf1c675)

Fradkin, A., E. Grewal, D. Holtz, and M. Pearson, (2014). Bias and Reciprocity in Online Reviews: Evidence from Field Experiments on Airbnb. *Working Paper*.

Fradkin, A, (2015). Search Frictions and the Design of Online Marketplaces. *Working Paper*.

Fryer, R., and S. Levitt. (2004). The Causes and Consequences of Distinctively Black Names. *The Quarterly Journal of Economics*, 119(3), 767-805.

Ghayad, R. (2014). *The Jobless Trap*. *Working Paper*.

Goldin, C., & Rouse, C. (2000). Orchestrating Impartiality. *American Economic Review*, 90(4), 715–741.

Hanson, A., & Hawley, Z. (2011). Do Landlords Discriminate in the Rental



Housing Market? Evidence from an Internet Field Experiment in U.S. Cities. *Journal of Urban Economics*, 70(2-3), 99-114.

Lahey, J. N. (2008). Age, Women, and Hiring: An Experimental Study. *Journal of Human Resources*, 43(1), 30–56. Lewis, G. (2011). Asymmetric Information, Adverse Selection and Online Disclosure: The Case of eBay Motors. *American Economic Review*, 101(4): 1535-46.

Larson, E. & Harris, A. (2016). Airbnb Sued, Accused of Ignoring Hosts' Race Discrimination. *Bloomberg News*. Available at: <http://www.bloomberg.com/news/articles/2016-05-18/airbnb-sued-over-host-s-alleged-discrimination-against-black-man>

Lewis, G. (2011). Asymmetric Information, Adverse Selection and Online Disclosure: The Case of eBay Motors. *American Economic Review*, 101(4): 1535-46.

List, J.A. (2004). The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field. *The Quarterly Journal of Economics*, 119(1): 49-89.

Luca, M. forthcoming. User-generated Content and Social Media. *Handbook of Media Economics*.

Morton, F., F. Zettelmeyer, and J. Silva-Risso. (2003). Consumer Information and Discrimination: Does the Internet Affect the Pricing of New Cars to Women and Minorities? *Quantitative Marketing and Economics*, 1(1), 65-92.

Ondrich, J., Stricker, A., and Yinger, J. (1999). Do Landlords Discriminate? The Incidence and Causes of Racial Discrimination in Rental Housing Markets. *Journal of Housing Economics*, 8, 185-204.

Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market?

A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy*, 3(4), 148–171.

Pager, D. (2003). The Mark of a Criminal Record. *American Journal of Sociology*, 108(5), 937–975.

Pope, D. G., & Sydnor, J. R. (2011). What's in a Picture?: Evidence of Discrimination from Prosper.com. *Journal of Human Resources*, 46(1), 53–92.

Schatz, B., Feinstein, D., & Warren, E. (2016). Letter to the Edith Ramirez, Chairwoman of the Federal Trade Commission.

Todisco, M. (2015). Share and Share Alike? Considering Racial Discrimination in the Nascent Room-Sharing Economy. *Stanford Law Review Online*, 67, 121–129.

U.S. Department of Housing and Urban Development. (2013). Housing Discrimination Against Racial and Ethnic Minorities 2012.

Yinger, J. (1998). Evidence on Discrimination in Consumer Markets. *Journal of Economic Perspectives*, 12(2), 23–40.

Zhao, B., Ondrich, J., and Yinger, J. (2005). Why Do Real Estate Brokers Continue to Discriminate? Evidence from the 2000 Housing Discrimination Study. *Center for Policy Research*. Paper 96.

## **Chapter 3: The Effect of Graphic Warning Labels on Sugary Drink Purchasing**

## **1. The Effect of Graphic Warning Labels on Sugary Drink Purchasing<sup>1</sup>**

Consumption of sugary drinks, such as soda, is a leading contributor to major health problems including obesity (Ludwig, Peterson, & Gortmaker, 2001), diabetes (Schulze et al., 2004), and heart disease (Fung et al., 2009). To reduce purchasing and consumption of sugary drinks, several local and state governments have proposed warning labels highlighting health risks; for example, San Francisco passed a policy requiring text warning labels on sugary drink advertisements, but it has not been implemented due to legal challenges from industry (Wiener, Mar, Cohen, & Avalos, 2015). Despite these initiatives, there are no published field tests evaluating whether sugary drink warning labels achieve their intended purpose in the real world, though two recent scenario-based lab studies point to their promise (Roberto, Wong, Musicus, & Hammond, 2016; VanEpps & Roberto, 2016). Beyond the question of effectiveness, there have been no published nationally-representative polls evaluating whether the public would accept them.

Like calorie labels, warning labels aim to provide health-relevant information to induce healthy behavior change. However, past research underscores the limits of this approach; for example, the evidence on whether calorie labels reduce calorie purchasing is mixed (Bleich et al., 2017; Downs, Wisdom, Wansink, & Loewenstein, 2013). Unlike calorie labels, warning labels convey direct information about potential health harms, which might increase their potency. This information attempts to overcome factors that often lead to suboptimal health decisions, including visceral factors such as hunger and self-control limits (FDA, 2014; Loewenstein, Read, & Baumeister, 2003; Stroebe, van Koningsbruggen, Papies, & Aarts, 2013).

---

<sup>1</sup> Co-authored with – in alphabetical order – Grant Donnelly, Leslie John, and Laura Zatz.

In addition to identifying the limitations of health information, research also suggests that it can be effective when provided in a salient and intuitively comprehensible way (Downs, Loewenstein, & Wisdom, 2009; Fagerlin, Zikmund-Fisher, & Ubel, 2011; Korfage et al., 2013). For example, people choose healthier beverages when calories are expressed in physical activity equivalents (Bleich, Barry, Gary-Webb, & Herring, 2014; Bleich, Herring, Flagg, & Gary-Webb, 2012). Such translation is likely even more compelling when it triggers an affective response (Loewenstein, 1996).

Indeed smoking cessation research has shown that graphic warning labels can be more effective than text warnings across a variety of outcomes (Noar et al., 2017; Noar et al., 2016a; Noar et al., 2016b; Purmehdi, Legoux, Carrillat, & Senecal, 2017). Sometimes a “diminishing cascade of effects” (Purmehdi et al., 2017) is observed whereby effects are strongest and most consistent for proximal measures of affective arousal and attention, less so for behavioral intentions, and weakest for behaviors such as calls to quit lines and cigarette consumption. Nonetheless even modest effects are noteworthy for such an intractable behavior given that labeling is a relatively weak intervention compared to approaches such as taxation and choice architecture.

Given the significant influence of public opinion on policy (Burstein, 2003) we also assessed consumer sentiment for placing graphic warning labels on sugary drinks. A recent study found that Americans generally prefer interventions that invoke primarily cognitive processes (e.g., facts about smoking risks) over those that invoke affect (e.g., pictures of cancer patients); however, support for the latter increased when people were informed of their effectiveness (Sunstein, 2016). We hypothesized that support for graphic warnings could be improved by conveying effectiveness information.

In sum, we: field-tested the effectiveness of graphic warning labels versus text, calorie, and no labels, and in a separate consumer survey, assessed consumer sentiment.

## **2. Methods**

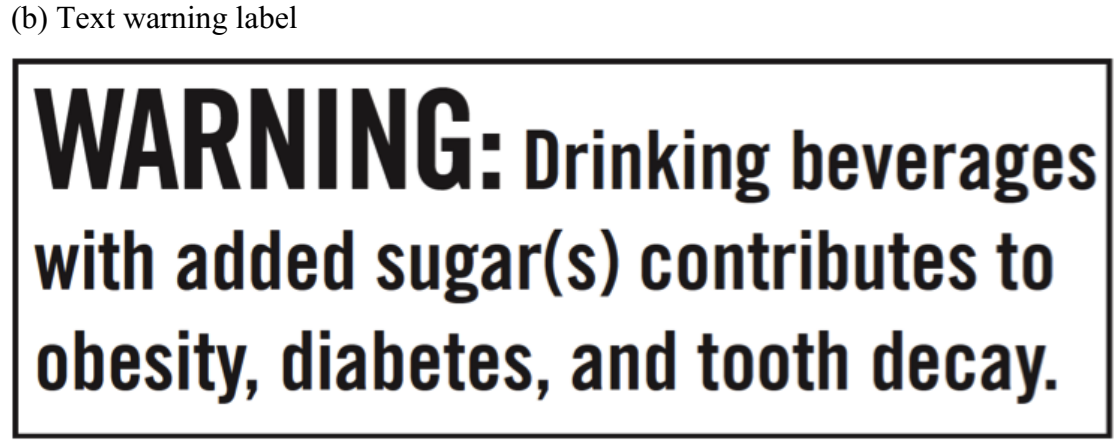
The field study occurred in a hospital cafeteria in Massachusetts over 14 weeks (April-July 2016) beginning with a two-week baseline to collect beverage sales data. Next, each sugary drink labeling intervention ran for two weeks, each followed by a two-week washout period when no labels were displayed (cf. Bleich et al., 2014). We pre-specified that each intervention would run for two weeks, based on a power analysis with the following parameters: 95% power ( $\beta = 0.05$ ), Type I error rate of 5% ( $\alpha = 0.05$ ), a small effect (Cohen's  $d = 0.20$ ), pre-baseline sales of 378 sugary drinks and 1,721 non-sugary drinks sold per week (based on one month of sales data of bottled drinks from February 2016, two months prior to our baseline period), and assuming a Fisher's exact statistical testing procedure. This power analysis suggested that we would only need to test each label for one week. However to be conservative, a priori, we decided to use two-week intervals. Table 3.1 depicts the study timeline. The study was preregistered at ClinicalTrials.gov (NCT02744859). Stimuli and data for this and both subsequent studies are available here: <https://osf.io/rh8pv/>.

Table 3.1: Study Timeline

| Baseline       |                  |                      | Intervention    |                      |                    | Post-<br>intervention |
|----------------|------------------|----------------------|-----------------|----------------------|--------------------|-----------------------|
| <u>2 weeks</u> | <u>2 weeks</u>   | <u>2 weeks</u>       | <u>2 weeks</u>  | <u>2 weeks</u>       | <u>2 weeks</u>     | <u>2 weeks</u>        |
| No label       | Calorie<br>label | Washout:<br>No label | Text<br>warning | Washout:<br>No label | Graphic<br>warning | No label              |

The hospital defined sugary drinks as any beverage with more than 12 grams of sugar per container (excluding milk and 100% juice). Drinks not meeting these criteria were not labeled. The calorie label followed a U.S. Food and Drug Administration regulation and read: “120–290 calories per container. 2,000 calories a day is used for general nutrition advice, but calorie needs vary” (FDA, 2016). The text warning label used the language proposed in San Francisco: “WARNING: Drinking beverages with added sugar(s) contributes to obesity, diabetes, and tooth decay.” The graphic warning label included the same text as the text warning label, but also included images portraying obesity, diabetes, and tooth decay (Figure 1). We chose images that were similarly evocative to those found to be effective on tobacco products.

**Figure 1.** Labels used: (a) calorie label; (b) text warning label; and (c) graphic warning label.



All bottled sugary drinks were grouped. On the cooler shelves immediately below the sugary drinks, we placed 12 salient 8 x 3 inch labels with large font (Appendix Figures D.1 & D.2). For fountain drinks, a 2½ x 1¼ inch label was displayed on each sugary drink dispenser (Appendix Figure D.3), for a total of four labels on the fountain machine. To minimize concerns



that the labels could shift people toward buying sugary drinks elsewhere, as opposed to truly decreasing sugary drink purchasing, we also displayed labels in front of sugary drinks in the building's alternate retail outlet (five labels) and vending machines (five labels); however, we did not collect their sales data.

Our primary interest was whether the labels shifted consumers away from purchasing sugary drinks. Therefore, and consistent with recent soda labeling research (e.g., VanEpps & Roberto, 2016; Bleich et al., 2012), our primary outcome measure was the proportion of drinks purchased that were sugary drinks.<sup>2</sup> This measure was superior to absolute units of sugary drinks purchased because it was less susceptible to sales fluctuations irrelevant to our treatment, such as differences in purchasing due to the day of the week (i.e., weekday vs. weekend) or seasonality; as a result we deemed the proportion measure to be both more valid and less noisy than the unit measure. For example, if the number of customers in the cafeteria doubled during one week, but the customers were drawn from the same population in terms of drink preferences, then the absolute units of sugary drinks purchased would double; however, one should not infer from this that the treatment that happened to be in place that week led individual customers to double their sugary drink purchasing. The change in absolute purchasing could therefore mask the variable of most interest: customers' drink choices.

This logic was supported by bottled drink sales data from February 2016 that we analyzed prior to the start of our study to inform its design. This analysis indicated that the number of drinks purchased varied across days (namely, decreasing on weekend days and

---

<sup>2</sup> When we first attempted to pre-register the proportion measure on ClinicalTrials.gov, the administrator rejected the measure, providing feedback which we interpreted as mandating that the outcome be a raw measure (i.e., absolute unit sales) as opposed to a transformation (e.g., proportion sales). After changing the outcome to a raw measure (i.e., number of beverages) the pre-registration was accepted. For transparency and to allow readers to assess the robustness of our findings, we report results using both the proportion and the unit measures.

holidays), whereas the proportion of sugary drinks purchased remained relatively constant, such that absolute units purchased mainly reflected changes in the number of customers rather than changes in drink choices. Nonetheless for completeness, we also provide the results using the units of sugary drinks purchased.

For secondary outcomes, we assessed beverage calories purchased, overall beverage purchases, share of drink types purchased, and the weight of fountain syrup dispensed.

Two data sources were used to measure outcomes due to differences in how the cafeteria's point-of-sale system recorded beverage purchases. For bottled drinks, each specific type (i.e., unique size and flavor) of bottled drink had its own product code; however, for fountain drinks, the system only recorded beverage size (not type or flavor). Thus, for bottled drinks, the data source was the cafeteria's point-of-sales system, which provided a daily summary of the number of each beverage type that had been purchased. For fountain drinks, a researcher weighed the boxes of syrup that were mixed with carbonated water to produce fountain drinks; the weights were recorded once per week. Each type (e.g., Diet Coke) used a unique ratio of carbonated water to syrup when dispensing a drink, so the weight of syrup was converted to a proxy for the number of fountain drinks sold by type.

### **3. Results**

During the study period, an average of 2,548 ( $SD = 290.0$ ) bottled drinks were purchased weekly ( $NS$  between weeks), approximately 20.5% ( $SD = 1.6\%$ ) of which were sugary drinks. Below, we report the results of analyses for our primary outcome of the proportion of sugary drinks purchased, followed by the same analyses using the units of sugary drinks purchased as the outcome. We then report the results for the number of calories purchased. Next, we ran a substitution analysis in which we assessed whether the label caused people to buy other types of

drinks, or to forego purchasing a drink altogether. Lastly, we examined whether the effect of the label was constant throughout the two week period in which it was present.

### **3.1. Proportion of sugary drinks purchased**

Our primary analysis was a Fisher's exact test of the proportion of bottled sugary drinks purchased by treatment. This was the simplest and most powerful statistical test of our interventions on sugary drink purchasing, and the test on which our power analysis was based. During baseline, 21.4% of bottled drinks purchased were sugary drinks. This percent was statistically indistinguishable from the share of sugary drinks purchased during the calorie label (21.5%, Fisher's exact  $p = .84$ ) and text warning label interventions (21.0%, Fisher's exact  $p = .66$ ). However, during the graphic warning label intervention, the average daily share of sugary drinks purchased decreased to 18.2% (Fisher's exact  $p < .001$ ), for an overall drop of 3.2 percentage points (which is a 14.8% reduction compared to baseline consumption). Graphic warning labels also reduced purchasing relative to both calorie (Fisher's exact  $p < .001$ ) and text warning labels (Fisher's exact  $p = .001$ ).

Next, because we tested our labels consecutively as opposed to concurrently, we considered possible effects of seasonality in two ways. This was important because seasonal changes in drinking habits over the course of our study could potentially confound the relationship between the labels and beverage sales. We first calculated descriptive statistics for the proportion of bottled sugary drinks purchased during each of our intervention periods as well as each two-week calendar period from 2014 and 2015 that matched our intervention periods (Table 3.2).

During 2016, when the graphic warning labels were displayed, there was a drop in the proportion of sugary drinks purchased – a drop that did not occur during the same calendar

period in either of the prior two years. Thus the descriptive statistics provided preliminary evidence that the decreased purchasing during the graphic warning label treatment was not a byproduct of cyclical sales.

Table 3.2: Bottled drink purchases during field study intervention periods (2016) and matched historical control periods (2014, 2015)

| 2016                            |                      |                          | 2015           |                      |                          | 2014           |                      |                          |
|---------------------------------|----------------------|--------------------------|----------------|----------------------|--------------------------|----------------|----------------------|--------------------------|
| Calendar Dates                  | Total Bottles Bought | Sugary Drinks Bought (%) | Calendar Dates | Total Bottles Bought | Sugary Drinks Bought (%) | Calendar Dates | Total Bottles Bought | Sugary Drinks Bought (%) |
| 4/25 – 5/8<br>(Baseline)        | 5,085                | 1087<br>(21.4)           | 4/27 – 5/10    | 5,359                | 1018<br>(19.0)           | 4/28 – 5/11    | 4,420                | 977<br>(22.1)            |
| 5/9 – 5/22<br>(Calorie Labels)  | 5,414                | 1166<br>(21.5)           | 5/11 – 5/24    | 6,816                | 1377<br>(20.2)           | 5/12 – 5/25    | 4,721                | 958<br>(20.3)            |
| 6/6 – 6/19<br>(Text Warning)    | 4,863                | 1021<br>(21.0)           | 6/8 – 6/21     | 5,865                | 1126<br>(19.2)           | 6/9 – 6/22     | 4,954                | 991<br>(20.0)            |
| 7/4 – 7/17<br>(Graphic Warning) | 5,021                | 914<br>(18.2)            | 7/6 – 7/19     | 5,362                | 1206<br>(22.5)           | 7/7 – 7/14     | 4,491                | 1010<br>(22.5)           |

Second, we conducted a series of regression analyses to test whether the results held when controlling for seasonality. We started with an unadjusted multivariable regression to predict the proportion of bottled drinks purchased that were sugary drinks on each day of our study, with dichotomous independent variables for each of the three labeling interventions (Appendix Table D.1). We used a robust variance estimator to account for heteroskedasticity.

The reference category was the baseline period, so coefficients on each of the dichotomous independent variables indicated differences relative to baseline. We then sequentially added seasonality covariates. To test whether two labeling interventions differed from each other, we ran the unadjusted regression, but changed the reference category to the intervention period of interest (i.e., the calorie label period would be the reference period when comparing the graphic warning label period to the calorie label period).

In the unadjusted model (Appendix Table D.1, model 1), the daily proportion of sugary drinks purchased was 3.4 percentage points lower during the graphic warning label period compared to baseline,  $\beta = -0.034$ ,  $SE = 0.01$ ,  $p = .001$ , but it was constant during the calorie and text warning labels.

When controlling for historical sales by adding fixed calendar week effects (i.e., the average proportion of sugary drinks sold in the same calendar week in 2014 and 2015), the proportion of sugary drinks purchased was constant during the calorie and text warning labels, but declined by 5.9 percentage points during the graphic warning label treatment,  $\beta = -0.059$ ,  $SE = 0.023$ ,  $p = .01$  (Appendix Table D.1, model 2). In other words, the effect of the graphic warning labels became stronger when controlling for historical sales. The effect was also robust to the addition of a control for heat index (calculated using the mean daily temperature and mean daily humidity). In this model, the daily proportion of sugary drinks purchased declined by 6.3 percentage points,  $\beta = -0.063$ ,  $SE = 0.022$ ,  $p < .001$ ; the coefficient for heat index was not statistically significant (Appendix Table D.1, model 3).

Although our analyses are focused on bottled drinks, results of a parallel analysis for fountain drinks also revealed a statistically significant effect of graphic warning labels on sugary drink purchasing (Supplement D). We focused on bottled drinks for several reasons. First, the

vast majority (about 90%) of drink purchases were bottled drinks. Second, focusing on bottled drinks enabled us to control for seasonality, which was not possible for fountain drinks because: a) we did not have historical data on changes in fountain syrup weight so we were unable to control for fixed calendar week effects; and b) fountain drink data were measured at the weekly level which would limit the number of observations per treatment in the regressions to two and prevent us from controlling for daily heat index. Therefore, the fountain drink analysis was restricted to the Fisher's exact test. Third, sales data for the two drink formats (fountain versus bottled) were obtained from different data sources: change in syrup weight vs. number of units sold.

### **3.2. Units of sugary drinks purchased**

The results of analyses using the units of bottled sugary drinks purchased as the outcome were generally consistent, though weaker, than the results reported above which used the proportion of sugary drinks purchased.

The results of the primary analysis using Fisher's exact test were equivalent when using the units of bottled sugary drinks purchased as the outcome measure. During the graphic warning label period, consumers purchased fewer bottled sugary drinks compared to baseline (Fisher's exact  $p = .005$ ; Table 3.2). There was no significant difference between the baseline period and the calorie label period (Fisher's exact  $p = .25$ ) or between the baseline period and the text warning label period (Fisher's exact  $p = .31$ ).

The analyses to examine potential effects of seasonality on the units of bottled sugary drinks, are presented in Table 3.2 and Appendix Table D.2. Consistent with the proportion measure, the units of sugary drinks purchased declined when the graphic warning labels were displayed in 2016, but not during the same calendar period in 2014 or 2015 (Table 3.2). In all

regression models, the number of bottled sugary drinks purchased dropped during the graphic warning treatment by 10 to 20 bottles per day; however, this effect was not always statistically significant. We suspect this is because the unit sales outcome was much noisier than the proportion measure: the standard deviation for the absolute units of sugary drinks purchased during our study (38.7 bottles) was roughly half of the mean, whereas the standard deviation for the proportion measure (0.037) was 16% of the mean. Empirically, this noise occurred in large part because fewer customers frequented the cafeteria on weekend days and holidays: the number of sugary drinks purchased declined from nearly 100 bottles per day during weekdays to 25 bottles per day on weekends. Hence, the estimated drop in units of sugary drinks purchased had a much wider confidence interval when holidays and weekend days were not controlled for in the regression. This phenomenon was not an issue for our primary results using the proportion measure because the proportion of sugary drinks sold was similar on holidays/weekend days and weekdays.

In the unadjusted regression model, there was not a statistically significant decline in the units of sugary drinks sold during the graphic warning label intervention,  $\beta = -12.36$ ,  $SE = 13.77$ ,  $p = .37$  (Appendix Table D.2, model 1). Controlling for holiday and weekend effects substantially reduced error variance, though the graphic warning label treatment did not reach statistical significance under this specification,  $\beta = -12.36$ ,  $SE = 7.01$ ,  $p = .08$  (Appendix Table D.2, model 2); the coefficient for the holiday and weekend effects was statistically significant,  $\beta = -69.70$ ,  $SE = 3.45$ ,  $p < .001$ . When we controlled for historical sales, further reducing error variance, the graphic warning label treatment was significantly different from baseline,  $\beta = -19.45$ ,  $SE = 9.39$ ,  $p = .044$  (Appendix Table D.2, model 3). When adding the heat index control, the effect was in the predicted direction, but the difference was not statistically significant,  $\beta = -$

19.77,  $SE = 10.96$ ,  $p = .078$  (Appendix Table D.2, model 4); the heat index covariate was not significant.

For fountain drinks, the effect of graphic warning labels on units of sugary drinks purchased did not reach statistical significance (Fisher's exact  $p = .52$ ) as it did for the proportion measure.

### **3.3. Beverage calories purchased**

To assess the impact of the labels on beverage calories purchased, we conducted a multivariable regression analysis in which the dependent variable was the average calories per bottled drink purchased in a given day during our treatment, with dichotomous independent variables for each of the three label interventions. We used a robust variance estimator to account for heteroskedasticity.

At baseline, the average calories per bottled drink purchased was 88 calories, 95% CI = [83 calories to 93 calories]; during the graphic warning label treatment, this average declined to 75 calories, 95% CI = [71 calories to 78 calories],  $p < .001$  (Appendix Table D.2). There was no statistically significant decline in calories per drink purchased during the calorie label treatment or text warning label treatment; the average calories per drink purchased was 86 calories, 95% CI = [81 calories to 90 calories],  $p = .58$ , and 85 calories, 95% CI = [81 calories to 89 calories],  $p = .47$ , respectively.

### **3.4. Substitution**

To assess substitution effects, we ran two analyses. First, to determine whether the labels caused people to refrain from buying drinks, we ran a multivariable linear regression in which the dependent variable was total bottled drinks purchased, with dichotomous independent variables for each of the three label interventions. We used a robust variance estimator to account



for heteroskedasticity. The unit of observation was one day. There were no significant differences in overall bottled drink sales by treatment.

Next, for any labels that reduced sugary drink purchases, we assessed whether, within bottled drink purchases, participants switched from sugary drinks to other types of drinks. We divided bottled drinks into four categories: water (including zero calorie sparkling and zero calorie flavored), non-sugary drinks with fewer than 20 calories (diet drinks), non-sugary drinks with at least 20 calories (e.g., unflavored milk), and sugary drinks. We ran a similar regression as above, but the dependent variable was the share of bottled drinks purchased corresponding to a given category. The daily proportion of water drinks purchased increased during the graphic warning intervention, from 24.9% at baseline to 28.1%,  $\beta = 0.032$ ,  $SE = 0.001$ ,  $p < .001$ , while purchasing of other drink types was unchanged. Therefore it seems that graphic warning labels led some consumers to buy water in lieu of sugary drinks.

### **3.5. Duration of Treatment Effect**

Lastly, we considered how the effectiveness of a label might change over time, both while in effect and once removed. We conducted an exploratory analysis that plotted the daily proportion of sugary drinks purchased and examined it for discernible patterns (Appendix Figure D.4). There was no discernible pattern, suggesting that label impact did not change throughout the two-week intervention periods. Notably, during the graphic warning label intervention – the only intervention that was effective – the decrease in sugary drink purchasing was observed consistently throughout the two-week period. In other words, it was not the case that a large immediate effect dissipated over the two-week period. After removing these graphic labels, sugary drink purchases rebounded to baseline levels. Specifically, the average daily proportion of drinks purchased that were sugary drinks was 21.9% during baseline, 18.5% in the graphic

warning label intervention, and 21.6% in the two-week period following this intervention, a significant rebound ( $p = .01$ ).

#### **4. Nationally Representative Survey**

##### **4.1. Survey Methods**

This study assessed public sentiment toward graphic warning labels, comparing it to two relevant benchmarks: calorie labels, a policy that has been implemented in several U.S. cities and states; and text warnings, a policy currently being appealed in San Francisco. Relative to these benchmarks, we expected support for graphic warning labels to be lower; however, we hypothesized that support for graphic warnings would be increased when conveying effectiveness information (i.e., the results of the field test).

We conducted a nationally representative online survey with participants ( $N = 402$ ; 49.8% female; *median* age = 45-54 years; 74.6% Caucasian; *median* annual income = \$25,000 - \$49,999; 55.5% attended at least some college; 28.3% consumed at least 1 sugary drink per day;  $M_{\text{BMI}} = 28.51$ ,  $SD = 7.57$  (excluding missing or implausible values); all *NS* between conditions) recruited through a survey company. The company obtains nationally representative samples by taking the pre-specified sample size and determining the required quotas for demographic variables (i.e., age, gender, ethnicity, Hispanic, income, education), and then recruits based on these quotas. We pre-specified a sample size of 400 based on both current suggested guidelines for sample size in behavioral research (Simmons, 2014; Simmons et al., 2011) and a power analysis using the parameters: power set to 90% ( $\beta = 0.10$ ), Type I error rate of 5% ( $\alpha = 0.05$ ), and a small effect (Cohen's  $d = 0.10$ ).

Participants viewed the three labels from the field study in a counterbalanced order. For each, they answered: "Do you support putting this label on sugar-sweetened beverages?" on a 7-

point scale from 1 (*Strongly Oppose*) to 7 (*Strongly Support*) (VanEpps & Roberto, 2016). Half of participants were randomly assigned to see effectiveness information accompanying the label. Specifically, for the calorie and text warning labels, participants were told that a recent study found that the label did not affect sugary drink purchasing. For the graphic warning label, participants were told that a recent study found the label reduced sugary drink purchasing and were informed of the magnitude of this effect.

Prior to running this study with a nationally representative sample, we conducted a pilot version using a large convenience sample and obtained the same result as that reported below (Supplement D). In addition, in the pilot study, we manipulated whether participants rated only one label versus all three. The results did not depend on this factor; therefore to reduce costs for the main, nationally representative survey, each participant rated the three labels, with the order counterbalanced between-subjects. In the main study reported here, there were no order effects; hence the reported results collapse across order.

The study was preregistered at ClinicalTrials.gov (NCT02947802).

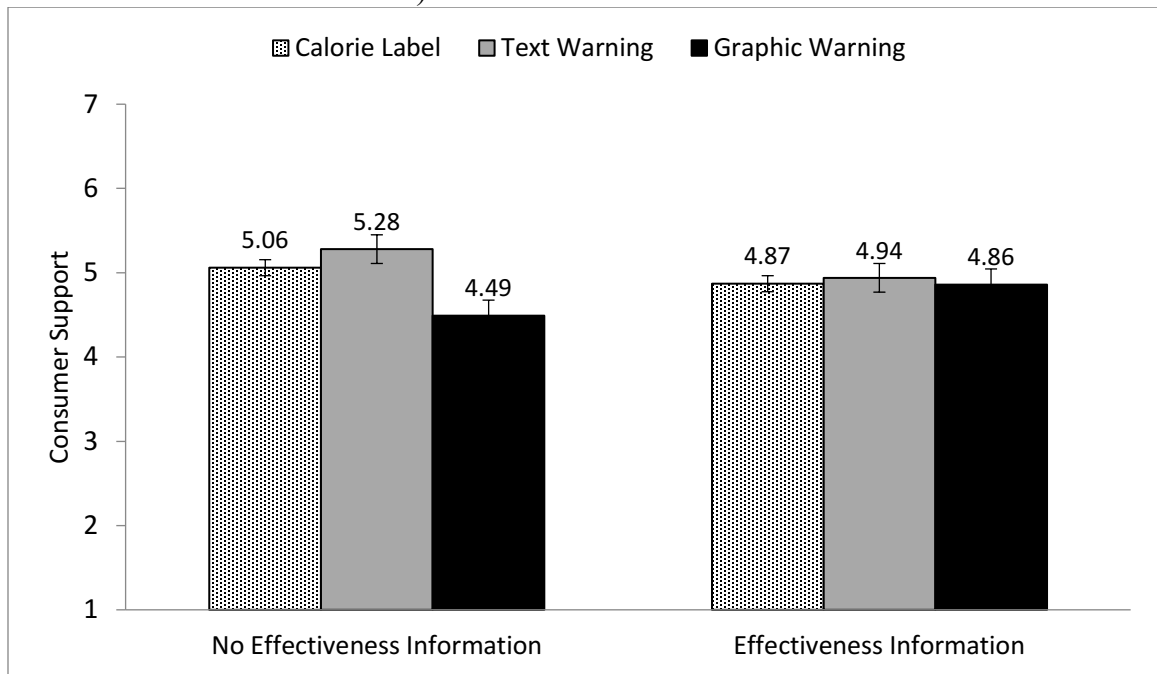
## **4.2. Survey Results**

A repeated-measures ANOVA using label type as a within-subjects factor and effectiveness information as a between-subjects factor revealed a significant main effect of label type,  $F(1.67, 669.26) = 12.06, p < .001$ , which was qualified by a significant interaction,  $F(1.67, 669.26) = 8.45, p = .001$  (Figure 3). (Mauchly's test indicated that the sphericity assumption was violated; therefore we use Greenhouse-Geisser estimates). Follow-up tests indicated that in the absence of effectiveness information, support for graphic warnings was significantly lower than both calorie labels,  $t(201) = -3.80, p < .001, d = 0.53$ , and text warnings,  $t(201) = -6.31, p < .001, d = 0.89$ . However, this effect was buffered by the provision of effectiveness information.

Specifically, when effectiveness information was given, support of graphic warnings was equivalent to both calorie labels,  $t(199) = -0.07, p = .95, d = 0.01$ , and text warnings,  $t(199) = -0.62, p = .54, d = 0.09$ .

Perhaps a more intuitive way of characterizing the results is to compare the percent of participants indicating support for the given label (i.e., responded above the neutral midpoint of the 7-point scale) across conditions. Consistent with the means reported above, in the absence of effectiveness information, a significantly smaller percent of participants supported the graphic warnings (50.8%) relative to both calorie labels (61.9%),  $z = 2.11, p = .03$ , and text warnings (66.8%),  $z = 3.14, p = .002$ . However, when effectiveness information was given, the percent of participants who supported the graphic warnings (55.6%) was statistically equivalent to both calorie labels (51.5%),  $z = 0.80, p = .42$ , and text warnings (56.5%),  $z = 0.20, p = .84$ .

**Figure 3.** Consumer support of labels by effectiveness information in national survey ( $N = 402$ ; error bars indicate  $\pm 1 SE$  of the mean)



Finally, an additional, exploratory analysis indicated that the effect of effectiveness information on label support did not depend on whether the given participant indicated that they

drink ( $N = 335$ ) versus do not drink ( $N = 67$ ) sugary drinks. Specifically, the 3-way interaction between label type (*calorie vs. text warning vs. graphic warning*), effectiveness information (*provided vs. not provided*), and sugary drinker status was not significant,  $F(1.67, 662.8) = .72, p = .46$ ; nor was the 2-way interaction between effectiveness information and sugary drinker status,  $F(1, 398) = .31, p = .58$ .

## **5. Discussion**

Our field study suggests that point-of-sale graphic warning labels reduced the proportion of sugary drinks purchased, driving people to buy water instead of sugary drinks, whereas calorie and text warning labels did not. Consistent with this pattern, when the graphic warning labels were removed, sugary drink purchasing rebounded to baseline levels. Our national survey suggests graphic warning labels are supported more if their effectiveness is conveyed. Although the observed increase in support for graphic warnings was small, support then matched that of the benchmark labels—namely calorie labels, which have been implemented in several jurisdictions. Interestingly, but generally consistent with Sunstein (2016), graphic label support did not surpass support for the other labels.

These findings offer guidance on providing information in a way that prompts healthier drink purchasing. While the text and graphic warning labels conveyed the same facts about health risks, only the more evocative graphic labels were associated with behavior change. Consistent with this finding, a recent lab study in New Zealand found that graphic warning labels decreased sugary drink purchase intentions (Bollard, Maubach, Walker, & Ni Mhurchu, 2016).

As the first field test of the effectiveness of graphic warning labels versus text warnings or calorie labels, our findings have legal implications. A federal attempt to mandate graphic warning labels for cigarettes failed in part due to a lack of field evidence proving that graphic

warnings were not “more extensive than necessary” (*R.J. Reynolds Tobacco Co. v. U.S. Food & Drug Administration*, 2012). Our findings may provide necessary evidence to implement graphic sugary drink warning labels.

Labeling, a form of information provision, is one of several strategies in policymakers’ “toolbox” to reduce sugary drink purchasing and intake; other strategies include pricing (i.e., taxes and subsidies) and choice architecture (i.e., structuring the environment to encourage better choices). How have these other approaches fared? Evaluations of sugary drink taxes are promising. A one peso per liter tax in Mexico led to a 5.5% decrease in the per capita volume of sugary drinks purchased in year one and 9.7% in year two (Colchero, Rivera-Dommarco, Popkin, & Ng, 2017). A one cent per ounce tax in Berkeley led to a 9.6% decrease in the volume of sugary drinks per transaction (Silver et al., 2017). As for choice architecture, reducing portion sizes can decrease consumption (Hollands et al., 2015; Rolls, Morris, & Roe, 2002), but implementation matters. For example, a portion cap like the one proposed by New York City could increase sugary drink purchasing when free refills are served (John, Donnelly, & Roberto, 2017). Future research might explore potential synergies in combining labeling, pricing, and choice architecture interventions.

This research is subject to several limitations. First, due to practical constraints, the field intervention ran consecutively in a single site. It is difficult to randomize individuals to different (but concurrent) interventions in a real-world cafeteria setting; this would introduce contamination issues and artificiality, threatening validity. We controlled for possible seasonality effects. Moreover, our design choice paralleled past field research which found that the order in which labels were tested did not matter (Bleich et al., 2012; Bleich et al., 2014). Nonetheless it is

possible that the effect of the graphic warning label was a product of the cumulative effect of previous labels.

Second, we could not assess how sugary drink purchasing might have changed outside the cafeteria. Customers might have foregone a sugary drink in the cafeteria only to buy one elsewhere. We minimized this possibility by posting the labels at the other locations where sugary drinks were sold in the building. Relatedly, we did not measure consumption. The calorie and text warning labels may not have been strong enough to reduce sugary drink purchasing, but they might have caused consumers to drink less of each container.

Future research could test the effect of label placement and design on purchasing and consumption. For example, warnings might be more effective when placed directly on beverage containers, where consumers would have repeated exposure as they drink; by contrast, point-of-sale warnings may be forgotten after purchasing. Interestingly, in contrast to the present investigation which found point-of-sale graphic warnings to be effective, two studies found that such warnings for tobacco did not affect purchasing (Kim et al., 2014; Coady et al., 2013); their warnings may have been less salient because they only used one large sign at the product display or one small sign at each register. With respect to design, we only tested one design for each label type, but many variations could be developed and tested.

Second, future research might also assess habituation in longer intervention periods. While the effect of the graphic warning label was consistent throughout the two-week period, tobacco warnings are more effective when their wording and design change over time (Borland et al., 2009; Wilson & Gilbert, 2008).

Future research might also investigate additional psychological processes underlying responses to warning labels. For example, do labels incite specific affective responses, such as

disgust or stigma? Graphic labels may introduce concerns over negative consequences such as “fat shaming.”

Fourth, future research could explore potential synergistic effects of calorie and warning labels, and whether effectiveness is influenced by how calorie information is presented. Understanding the effects of different types of calorie labels across settings is an ongoing area of inquiry; although not our primary focus, the field study also offers one data point for this discussion (Bleich et al., 2017; Block & Roberto, 2014).

Finally, studies could explore heterogeneity of effects (for example, by weight or socioeconomic status). Relatedly, although the labels in our field study were very salient, research could explore whether the labels are differentially noticed or persuasive by demographic characteristics. For example, individuals who are female, higher income, or health conscious are particularly attentive to calorie information (Bleich et al., 2017). To test the generalizability of our findings, this intervention could be tested in other retail settings with a large sample of diverse consumers. Our setting was a Northeast hospital where sugary drink purchasing was relatively low at baseline and information about calories or health risks may not have been novel for some consumers, which may have limited our ability to detect changes, particularly for calorie and text warning labels.

In conclusion, this research is the first test of the real-world effectiveness and acceptability of graphic sugary drink warning labels. Graphic warning labels decreased the proportion of sugary drinks purchased; significant changes were not observed for calorie labels or text warning labels. Consumer support for graphic warning labels can be increased by communicating their effectiveness. Taken together, these studies contribute to the psychology of healthy behavior change and provide evidence to inform policymakers.



## References

- Bleich, S. N., Barry, C. L., Gary-Webb, T. L., & Herring, B. J. (2014). Reducing sugarsweetened beverage consumption by providing caloric information: How Black adolescents alter their purchases and whether the effects persist. *American Journal of Public Health, 104*, 2417-2424. doi:10.2105/ajph.2014.302150
- Bleich, S. N., Economos, C. D., Spiker, M. L., Vercammen, K. A., VanEpps, E. M., Block, J. P., . . . Roberto, C. A. (2017). A systematic review of calorie labeling and modified calorie labeling interventions: Impact on consumer and restaurant behavior. *Obesity, 25*, 2018-2044. doi:10.1002/oby.21940
- Bleich, S. N., Herring, B. J., Flagg, D. D., & Gary-Webb, T. L. (2012). Reduction in purchases of sugar-sweetened beverages among low-income black adolescents after exposure to caloric information. *American Journal of Public Health, 102*, 329-335. doi:10.2105/ajph.2011.300350
- Block, J. P., & Roberto, C. A. (2014). Potential benefits of calorie labeling in restaurants. *Journal of the American Medical Association, 312*, 887-888. doi:10.1001/jama.2014.9239
- Bollard, T., Maubach, N., Walker, N., & Ni Mhurchu, C. (2016). Effects of plain packaging, warning labels, and taxes on young people's predicted sugar-sweetened beverage preferences: An experimental study. *International Journal of Behavioral Nutrition and Physical Activity, 13*, 95. doi:10.1186/s12966-016-0421-7
- Borland, R., Wilson, N., Fong, G. T., Hammond, D., Cummings, K. M., Yong, H.-H., . . . McNeill, A. (2009). Impact of graphic and text warnings on cigarette packs: Findings

- from four countries over five years. *Tobacco Control*, 18, 358-364.  
doi:10.1136/tc.2008.028043
- Burstein, P. (2003). The impact of public opinion on public policy: A review and an agenda. *Political Research Quarterly*, 56, 29-40. doi:10.1177/106591290305600103
- Colchero, M. A., Rivera-Dommarco, J., Popkin, B. M., & Ng, S. W. (2017). In Mexico, evidence of sustained consumer response two years after implementing a sugar-sweetened beverage tax. *Health Affairs*, 36, 564-571. doi:10.1377/hlthaff.2016.1231
- Downs, J. S., Loewenstein, G., & Wisdom, J. (2009). Strategies for promoting healthier food choices. *American Economic Review*, 99, 159-164. doi:10.1257/aer.99.2.159
- Downs, J. S., Wisdom, J., Wansink, B., & Loewenstein, G. (2013). Supplementing menu labeling with calorie recommendations to test for facilitation effects. *American Journal of Public Health*, 103, 1604-1609. doi:10.2105/ajph.2013.301218
- Emery, L. F., Romer, D., Sheerin, K. M., Jamieson, K. H., & Peters, E. (2014). Affective and cognitive mediators of the impact of cigarette warning labels. *Nicotine & Tobacco Research*, 16, 263-269. doi:10.1093/ntr/ntt124
- Evans, A. T., Peters, E., Strasser, A. A., Emery, L. F., Sheerin, K. M., & Romer, D. (2015). Graphic warning labels elicit affective and thoughtful responses from smokers: Results of a randomized clinical trial. *PLOS One*, 10, e0142879.  
doi:<https://doi.org/10.1371/journal.pone.0142879>
- Fagerlin, A., Zikmund-Fisher, B. J., & Ubel, P. A. (2011). Helping patients decide: Ten steps to better risk communication. *Journal of the National Cancer Institute*, 103, 1436-1443.  
doi:<https://doi.org/10.1093/jnci/djr318>

- FDA. (2014). Food labeling: Nutrition labeling of standard menu items in restaurants and similar retail food establishments. Final regulatory impact analysis. Retrieved from <https://www.fda.gov/downloads/Food/IngredientsPackagingLabeling/LabelingNutrition/UCM423985.pdf>
- FDA. (2016). A labeling guide for restaurants and retail establishments selling away-from-home foods—Part II (menu labeling requirements in accordance with the patient protection affordable care act of 2010). Retrieved from <https://www.fda.gov/downloads/Food/GuidanceRegulation/GuidanceDocumentsRegulatoryInformation/UCM461963.pdf>
- Fung, T. T., Malik, V., Rexrode, K. M., Manson, J. E., Willett, W. C., & Hu, F. B. (2009). Sweetened beverage consumption and risk of coronary heart disease in women. *The American Journal of Clinical Nutrition*, 89, 1037-1042. doi:10.3945/ajcn.2008.27140
- Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, 67, 451-470. doi:10.1111/bmsp.12028
- Hollands, G. J., Shemilt, I., Marteau, T. M., Jebb, S. A., Lewis, H. B., Wei, Y., . . . Ogilvie, D. (2015). Portion, package or tableware size for changing selection and consumption of food, alcohol and tobacco. *The Cochrane Database of Systematic Reviews*, CD011045. doi:10.1002/14651858.CD011045.pub2
- John, L. K., Donnelly, G. E., & Roberto, C. A. (2017). Psychologically informed implementations of sugary-drink portion limits. *Psychological Science*, 28, 620-629. doi:10.1177/0956797617692041

- Korfage, I. J., Fuhrel-Forbis, A., Ubel, P. A., Zikmund-Fisher, B. J., Greene, S. M., McClure, J. B., . . . Fagerlin, A. (2013). Informed choice about breast cancer prevention: Randomized controlled trial of an online decision aid intervention. *Breast Cancer Research*, 15, R74-R74. doi:10.1186/bcr3468
- Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65, 272-292.  
doi:https://doi.org/10.1006/obhd.1996.0028
- Loewenstein, G., Read, D., & Baumeister, R. F. (2003). Time and decision: Economic and psychological perspectives of intertemporal choice: Russell Sage Foundation.
- Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, 127, 267-286. doi:10.1037/0033-2909.127.2.267
- Ludwig, D. S., Peterson, K. E., & Gortmaker, S. L. (2001). Relation between consumption of sugar-sweetened drinks and childhood obesity: a prospective, observational analysis. *The Lancet*, 357(9255), 505-508. doi:https://doi.org/10.1016/S0140-6736(00)04041-1
- Noar, S. M., Francis, D. B., Bridges, C., Sontag, J. M., Brewer, N. T., & Ribisl, K. M. (2017). Effects of strengthening cigarette pack warnings on attention and message processing: A systematic review. *Journalism & Mass Communication Quarterly*, 94, 416-442.  
doi:10.1177/1077699016674188
- Noar, S. M., Francis, D. B., Bridges, C., Sontag, J. M., Ribisl, K. M., & Brewer, N. T. (2016a). The impact of strengthening cigarette pack warnings: Systematic review of longitudinal observational studies. *Social Science & Medicine*, 164, 118-129.  
doi:https://doi.org/10.1016/j.socscimed.2016.06.011

- Noar, S. M., Hall, M. G., Francis, D. B., Ribisl, K. M., Pepper, J. K., & Brewer, N. T. (2016b). Pictorial cigarette pack warnings: A meta-analysis of experimental studies. *Tobacco Control*, 25, 341-354. doi:10.1136/tobaccocontrol-2014-051978
- Peters, E. (2006). The functions of affect in the construction of preferences. In S. Lichtenstein & P. Slovic (Eds.), *The construction of preference* (pp. 454-463). New York: Cambridge University Press.
- Peters, E., Lipkus, I., & Diefenbach, M. A. (2006). The functions of affect in health communications and in the construction of health preferences. *Journal of Communication*, 56, S140-S162. doi:10.1111/j.1460-2466.2006.00287.x
- Peters, E., Romer, D., Slovic, P., Jamieson, K. H., Wharfield, L., Mertz, C. K., & Carpenter, S. M. (2007). The impact and acceptability of Canadian-style cigarette warning labels among U.S. smokers and nonsmokers. *Nicotine & Tobacco Research*, 9, 473-481. doi:10.1080/14622200701239639
- Purmehdi, M., Legoux, R., Carrillat, F., & Senecal, S. (2017). The effectiveness of warning labels for consumers: A meta-analytic investigation into their underlying process and contingencies. *Journal of Public Policy & Marketing*, 36, 36-53. doi:10.1509/jppm.14.047
- R.J. Reynolds Tobacco Co. v. U.S. Food & Drug Administration, (2012).
- Roberto, C. A., Wong, D., Musicus, A., & Hammond, D. (2016). The influence of sugar-sweetened beverage health warning labels on parents' choices. *Pediatrics*, 137, e20153185. doi:10.1542/peds.2015-3185

- Rolls, B. J., Morris, E. L., & Roe, L. S. (2002). Portion size of food affects energy intake in normal-weight and overweight men and women. *The American Journal of Clinical Nutrition*, 76, 1207-1213.
- Schulze, M. B., Manson, J. E., Ludwig, D. S., Graham, A. C., Meir, J. S., Walter, C. W., & Frank, B. H. (2004). Sugar-sweetened beverages, weight gain, and incidence of type 2 diabetes in young and middle-aged women. *Journal of the American Medical Association*, 292, 927-934. doi:10.1001/jama.292.8.927
- Silver, L. D., Ng, S. W., Ryan-Ibarra, S., Taillie, L. S., Induni, M., Miles, D. R., . . . Popkin, B. M. (2017). Changes in prices, sales, consumer spending, and beverage consumption one year after a tax on sugar-sweetened beverages in Berkeley, California, US: A before-and-after study. *PLOS Medicine*, 14, e1002283. doi:10.1371/journal.pmed.1002283
- Simmons, J. (2014). MTurk vs. the lab: Either way we need big samples, Data Colada.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359-1366. doi:10.1177/0956797611417632
- Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397-421). New York: Cambridge University Press.
- Stroebe, W., van Koningsbruggen, G. M., Papies, E. K., & Aarts, H. (2013). Why most dieters fail but some succeed: A goal conflict model of eating behavior. *Psychological Review*, 120, 110-138. doi:10.1037/a0030849
- Sunstein, C. R. (2016). People prefer system 2 nudges (kind of). *Duke Law Journal*, 66(121), 122-168.

- VanEpps, E. M., & Roberto, C. A. (2016). The influence of sugar-sweetened beverage warnings: A randomized trial of adolescents' choices and beliefs. *American Journal of Preventive Medicine*, 51, 664-672. doi:<https://doi.org/10.1016/j.amepre.2016.07.010>
- Wiener, S., Mar, E., Cohen, M., & Avalos, J. (2015). Sugar-sweetened beverage warning for advertisements San Francisco: San Francisco Board of Supervisors. Retrieved from <http://www.sfbos.org/ftp/uploadedfiles/bdsupvrs/ordinances15/o0100-15.pdf>.
- Wilson, T. D., & Gilbert, D. T. (2008). Explaining away: A model of affective adaptation. *Perspectives on Psychological Science*, 3, 370-386. doi:10.1111/j.1745-6924.2008.00085

**Appendix A:** Experimental instructions for internet privacy and information avoidance experiment

This section shows the instructions shown to each participant.

All participants start off by seeing the same introductory slide, shown in Figure A1.

Each group is then shown one of three possible instructions slides, shown in Figure A2. Participants then perform the task itself. Finally, after completing the privacy task associated with their treatment, all participants completed a short survey, shown in Figure A5. Participants who agreed to give up their Facebook data would see the “Log In with Facebook” button above the survey; participants who opted to remain anonymous would not see the button.

- You will fill out a short survey about your health and financial situation. Before doing the survey, you’ll make decisions about **the size of your bonus** and your **privacy settings**.
- Your **privacy settings** can be **anonymous**, or through **Facebook**.
- Either way you will do the same survey.
- If you choose the Facebook option, you will see a “Log in with Facebook” button above the survey. You will have to log in with your Facebook account. This means that the survey-taker will see your public Facebook profile, along with your survey answers.
- If you choose the anonymous setting, you will complete the survey anonymously.

Figure A1: This figure shows the introductory page of the instructions, which was shown to all groups.



- On the next screen, you'll see a table like this:
 

|  |              |
|--|--------------|
| <input type="radio"/> 2 cents                  | High Privacy |
| <input type="radio"/> 52 cents                 | Low Privacy  |
| <a href="#">Click here to make your choice</a> |              |
- The first column shows how big your bonus will be.
- The second column shows your privacy settings. "High Privacy" means doing the survey anonymously. "Low Privacy" means doing it after logging in through Facebook.
- Sometimes, the top row will be the high privacy option. Sometimes it will be the low privacy option. It can be either, with a 50/50 chance.

- On the next screen, you'll see a table like this:
 

|  |  |
|--|--|
| <a href="#">Click here to see the privacy settings</a> |  |
| <input type="radio"/> 2 cents                          |  |
| <input type="radio"/> 52 cents                         |  |
| <a href="#">Click here to make your choice</a>         |  |
- The first column shows how big your bonus will be.
- The second column – which you have to click to see – shows your privacy settings. "High Privacy" means doing the survey anonymously. "Low Privacy" means doing it after logging in through Facebook.
- Sometimes, the top row will be the high privacy option. Sometimes it will be the low privacy option. It can be either, with a 50/50 chance. You have to click to make sure.

- On the next screen, you'll see tables like this:
 

|  |              |
|--|--------------|
| <input type="radio"/> 2 cents                  | High Privacy |
| <input type="radio"/> 52 cents                 | Low Privacy  |
| <a href="#">Click here to make your choice</a> |              |
- The first column shows how big your bonus will be.
- The second column shows your privacy settings. "High Privacy" means doing the survey anonymously. "Low Privacy" means doing it after logging in through Facebook.
- Choose whichever option you prefer
- You'll face several choices like this. One of your choices will be randomly selected and enforced. So it always makes sense to just choose what you think you prefer.

- On the next screen, you'll see a table like this:
 

|  |  |
|--|--|
| <a href="#">Click here to see the second money bonus</a> |  |
| <input type="radio"/> 2 cents                            |  |
| <input type="radio"/> 52 cents                           |  |
| <a href="#">Click here to make your choice</a>           |  |
- The first column shows how big your first bonus will be.
- The second column – which you have to click to see – shows the second bonus. The size of the second bonus is \$X.XX.
- Sometimes, the top row will be the high bonus option. Sometimes it will be the low bonus option. It can be either, with a 50/50 chance. You have to click to make sure.

Figure A2: This figure shows the instructions page for each of the four treatments. The Direct Tradeoff Treatment group was shown the instructions in the top panel. The Veiled Tradeoff Treatment group was shown the instructions in the second panel. The Elicitation Treatment group was shown the instructions in the third panel. The Placebo Veiled Tradeoff group was shown the instructions in the bottom panel.

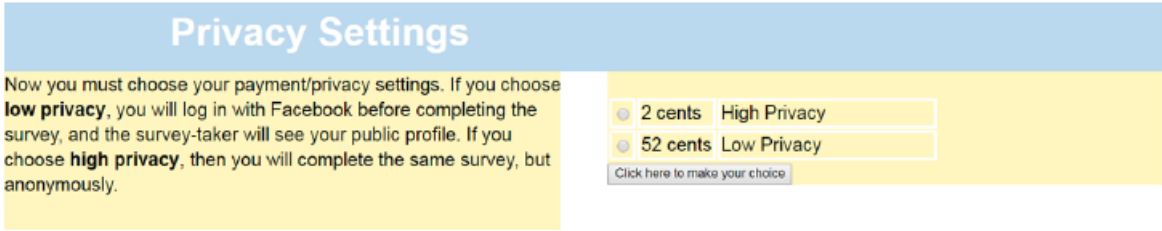


Figure A3: This is a screenshot of the privacy task page completed by participants in the Direct Tradeoff Treatment, after reading instructions and completing practice rounds. Participants in the Elicitation Treatment faced an identical task, but completed it multiple times, with monetary bonuses ranging from \$0.25 to \$5.02. As described in the instructions in the bottom panel of Figure A2, one of these choices would be randomly selected and enforced.

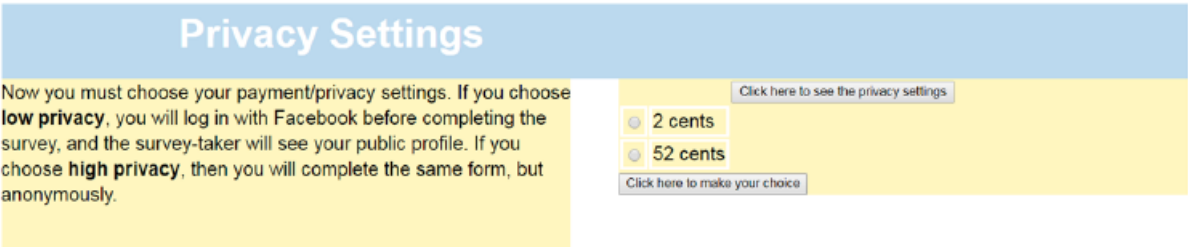


Figure A4: This is a screenshot of the privacy task page completed by participants in the Veiled Tradeoff Treatment, after reading instructions and completing practice rounds.



Age:

How many times do you work out in a typical week?

How many times have you tried to diet in your life?

What is your annual income range?

How much credit card debt do you have?

How carefully did you make your choices?

How old were you when you were 10 years old?

Do you have a Facebook account?

How often do you use Facebook in a typical week?

Figure A5: After making their privacy choices, all participants completed the survey above. Those participants who opted for the anonymous survey were not shown the Facebook login button. Those that opted for the low privacy setting saw the login button, as in the picture above.

## Appendix B: Model of Participants' Choice in Veiled Tradeoff Treatment

Consider the decision of whether a participant should click to reveal the privacy column. Let her utility from privacy be  $v$ , her utility from 50 cents be  $u(50)$ , and her utility from clicking be  $c$ .

If she does not click to reveal the privacy column, the participant will choose the 50 cent option, face a 50% chance of losing her privacy, and incur no clicking costs:

$$U_{\text{no click}} = u(50) + 0.5 \cdot v$$

Indeed, if  $v < u(50)$  (she values 50 cents more than privacy), there is no need to click to reveal. If she clicks and faces a tradeoff between money and privacy, she will choose the money anyway, so if  $v < u(50)$ , she will never click to reveal, since this will allow her to avoid clicking costs.

Suppose instead that  $v > u(50)$ , so the participant may want to click to reveal. The participant will click to reveal if the expected utility of clicking is larger than the expected utility of not clicking.

If she clicks to reveal, she will choose privacy with certainty, face a 50% chance of not getting the 50 cents, and incur clicking costs:

$$U_{\text{click}} = 0.5 \cdot u(50) + v + c$$

She will click if  $U_{\text{click}}$  is greater than the utility of not clicking:

$$\begin{aligned} 0.5 \cdot u(50) + v + c &> u(50) + 0.5 \cdot v \\ v - u(50) &> 2 \cdot c \end{aligned}$$

This demonstrates the second conclusion. If clicking costs are zero, a participant who values privacy more than 50 cents will always click to reveal. In general, she will click to reveal so long as the difference between her privacy valuation and the utility from 50 cents is larger than two times the clicking costs.

Given this decision rule, and given the distribution of WTP prices in the Elicitation Treatment, how big would clicking costs have to be to explain the treatment effect? If  $c = 0$ , then everyone with a WTP for privacy above 50 cents should click to reveal, leading to a click rate of 65%. If clicking costs are \$1.00, then anyone with a WTP for privacy of \$2.50 and above would click to reveal, leading to a click rate of roughly 50% – still higher than the observed click rate of 42%. Clicking costs would have to be nearly \$2.00 to end up with a click rate observed in the main experiment.

**Appendix C: Supplementary material from Airbnb Field Experiment Study**

**Appendix Table C.1: Results of survey testing races associated with names**

| <i>White Female</i> |      | <i>African-American Female</i> |      |
|---------------------|------|--------------------------------|------|
| Meredith O'Brien    | 0.93 | Tanisha Jackson                | 0.03 |
| Anne Murphy         | 0.95 | Lakisha Jones                  | 0.05 |
| Laurie Ryan         | 0.97 | Latoya Williams                | 0.05 |
| Allison Sullivan    | 0.98 | Latonya Robinson               | 0.07 |
| Kristen Sullivan    | 1.00 | Tamika Williams                | 0.07 |
| <i>White Male</i>   |      | <i>African-American Male</i>   |      |
| Greg O'Brien        | 0.88 | Tyrone Robinson                | 0.00 |
| Brent Baker         | 0.90 | Rasheed Jackson                | 0.06 |
| Brad Walsh          | 0.91 | Jamal Jones                    | 0.07 |
| Brett Walsh         | 0.93 | Darnell Jackson                | 0.10 |
| Todd McCarthy       | 0.98 | Jermaine Jones                 | 0.26 |

Notes: "White" is coded as 1. "African-American" is coded as 0. Sample size = 62.

**Appendix Table C.2: Discrimination by City**

*Dependent Variable: 1(Host Accepts)*

|  | All<br>Cities | Baltimore<br>(N = 347) | Dallas<br>(N = 415) | LA<br>(N = 3,913)  | St. Louis<br>(N = 151) | D.C.<br>(N = 1,559) |
|--|---------------|------------------------|---------------------|--------------------|------------------------|---------------------|
| Guest is Af-<br>Am   | -0.08***      | -0.07***<br>(0.02)     | -0.08***<br>(0.02)  | -0.10**<br>(0.02)  | -0.08***<br>(0.03)     | -0.08***<br>(0.02)  |
| City   | --            | 0.07<br>(0.03)         | 0.04<br>(0.03)      | -0.00<br>(0.03)    | 0.02<br>(0.05)         | -0.03<br>(0.04)     |
| City * Guest<br>is Af-Am   | --            | -0.12*<br>(0.05)       | -0.01<br>(0.04)     | 0.03<br>(0.04)     | 0.02<br>(0.07)         | -0.01<br>(0.05)     |
| Constant   | 0.49          | 0.48***<br>(0.01)      | 0.49***<br>(0.01)   | 0.49***<br>(0.02)  | 0.49***<br>(0.01)      | 0.50***<br>(0.01)   |
| Observations   | 6,235         | 6,235                  | 6,235               | 6,235              | 6,235                  | 6,235               |
| Adjusted R <sup>2</sup>  | 0.006         | 0.007                  | 0.006               | 0.006              | 0.006                  | 0.007               |
| Implied<br>Coef. on<br>Guest is Af-<br>Am + City *<br>Guest is Af-<br>Am | --            | -0.19***<br>(0.04)     | -0.09**<br>(0.04)   | -0.07***<br>(0.02) | -0.06<br>(0.06)        | -0.09*<br>(0.05)    |

Notes: Standard errors are clustered by (guest name)\*(city) and are reported in parentheses.

\* p < .10. \*\* p < .05. \*\*\* p < .01.

**Appendix Table C.3: Host responses to guest inquiries, by race of the guest**

|  | <i>White Guests</i> | <i>African-American<br/>Guests</i> |
|--|---------------------|------------------------------------|
| Yes                                    | 1,152               | 940                                |
| Yes, but request for more information  | 375                 | 308                                |
| Yes, with lower price if booked now    | 11                  | 10                                 |
| Yes, if guest extends stay             | 10                  | 15                                 |
| Yes, but in a different property       | 18                  | 8                                  |
| Yes, at a higher price                 | 4                   | 0                                  |
| Request for more information           | 339                 | 323                                |
| Not sure or check back later           | 154                 | 175                                |
| No response                            | 429                 | 423                                |
| No unless more information is provided | 12                  | 15                                 |
| No                                     | 663                 | 873                                |

Notes: The table reports the frequency of each type of host response to a guest inquiry, by race of the guest. Likelihood-ratio chi-squared = 68.61 ( $p < .01$ ). Null hypothesis is that the columns will have equal proportions for each type of response

## **Appendix D: Supplementary material from soda label study**

### **Fountain Drink Purchases During Field Study**

Our intervention included warning labels on a soda fountain machine, and we tested whether our results replicated for these drinks. For fountain drinks however, purchase data only included the size of the fountain cup purchased, not the flavor or type of beverage. To solve this problem, we measured changes in the amount of syrup used for each drink type by weighing the boxes of syrup once a week. Hence, if the box of Coca-Cola syrup saw a drop of 14 pounds, but the box of Diet Coke syrup saw a drop of 21 pounds, we could conclude that more Diet Coke syrup was dispensed.

Each drink used a unique ratio of water to syrup when dispensing a drink, written on the fountain machine itself. We used this ratio to convert the weight of syrup dispensed into number of fluid ounces dispensed. Finally, using data on number of fountain cups purchased, we divided the total number of fluid ounces by the average cup size purchased (21.8 ounces) to construct a proxy for the units of each drink that were purchased.

Fig D.6 shows the estimated proportion of sugary fountain drinks versus non-sugary fountain drinks purchased for the baseline period and each intervention period. We found the same results as for bottled beverage purchases. During the baseline period, 58% of the drinks purchased were sugary drinks. This was roughly unchanged during the calorie label intervention (57%,  $p = .76$ ) and during the text warning label intervention (54%,  $p = .23$ ). By contrast, the proportion of sugary drinks purchased dropped to 50% during the graphic warning labels intervention, a statistically significant drop when compared to baseline ( $p = .01$ ) and the calorie warning label intervention ( $p = .02$ ) but not the text warning label intervention ( $p = .20$ ). This



change during the graphic warning label period represents a 14% drop from baseline, almost precisely mirroring the drop in bottled sugary drinks purchased.

### **Consumer Support for Label: Pre-Test with Convenience Sample**

The nationally representative survey reported in the main text is a replication of pre-test which we conducted with a convenience sample ( $N = 254$ ; 44.1% female; 83.5% White). Specifically, as in the nationally representative sample, participants rated the extent to which they supported each label using the same scale. Participants were randomized to view and rate only one of the three labels (separate evaluation condition), or to view and rate all three (in which case order of presentation was randomized between-participants; joint evaluation condition). As in the nationally representative survey reported in the main text, for half of participants, effectiveness information accompanied the label (for the other half, effectiveness information was not provided).

**Joint evaluation.** A repeated-measures ANOVA using label type as a within-subjects factor and effectiveness information as a between-subjects factor revealed a significant main effect of label type,  $F(1.72, 106.33) = 6.08, p = .005$ , no effect for effectiveness information,  $F(1, 62) = 0.14, p = .71$ , but a significant interaction,  $F(1.72, 106.33) = 10.55, p < .001$ .<sup>1</sup> Follow-up tests revealed that when effectiveness information was provided, people were equally accepting of graphic warning labels relative to both calorie,  $t(30) = 1.92, p = .07$ , and text warning labels,  $t(30) = 1.03, p = .31$ . However, in the absence of such information, people were less accepting of graphic warning labels relative to text warning labels,  $t(32) = 5.65, p < .001$ , and equally accepting to calorie labels,  $t(32) = 1.69, p = .10$ . In sum, in the absence of

---

<sup>1</sup> Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of label type,  $\chi^2(2) = 11.09, p = .004$ . There was greater variance in support for the graphic label relative to the calorie and text warning label. Therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ( $\epsilon = .86$ ).

effectiveness information, consumers were neutral about graphic warning labels; however, this indifference can be turned into support by providing effectiveness information.

**Separate evaluation.** As noted, the other half of our sample evaluated only one label. A 2x2 ANOVA revealed a marginal main effect of label,  $F(2, 184) = 3.84, p = .02$ , as well as a significant main effect of effectiveness information,  $F(1, 184) = 3.97, p = .048$ . Importantly however, these main effects were qualified by a marginally significant interaction,  $F(2, 184) = 2.80, p = .06$ . Pairwise comparisons revealed that in the absence of effectiveness information, support for the graphic warning was lower relative to both the calorie label,  $t(56) = 2.05, p = .045$ , and marginally lower than the text warning label,  $t(64) = 1.70, p = .09$ . However, when effectiveness information was provided, respondents were just as supportive of the graphic warning as they were the calorie label,  $t(66) = -0.58, p = .63$ , although support was still significantly lower than the text warning,  $t(61) = 2.02, p = .048$ . These results are broadly consistent with those of the joint evaluation condition; therefore, in the main study to maximize power (and reduce costs, since the nationally-representative survey was conducted through a survey panel company which charged per respondent), all participants rated all three labels (i.e., we only ran the joint evaluation mode conditions).

## Supplemental Figures and Tables



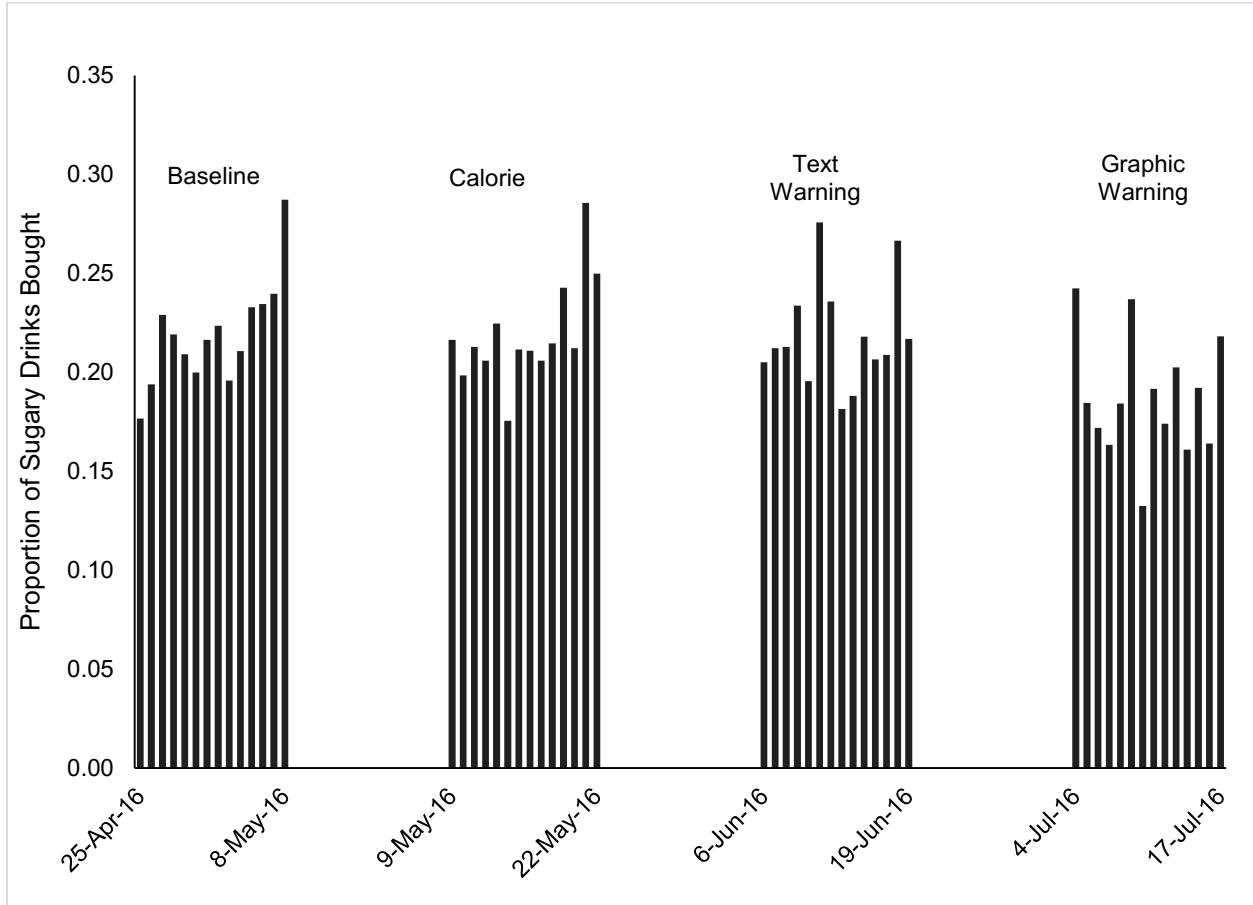
**Appendix Figure D.1.** Study 1: Bottled beverage cooler with sugary drinks on the top left during the calorie label intervention, and non-sugary drinks on the right and bottom shelves.



**Appendix Figure D.2.** Study 1: Bottled beverage cooler depicting the sugary drinks during the graphic warning label intervention.



**Appendix Figure D.3.** Study 1: Fountain drink machine depicting sugary drinks during the text warning label treatment and non-sugary drinks.

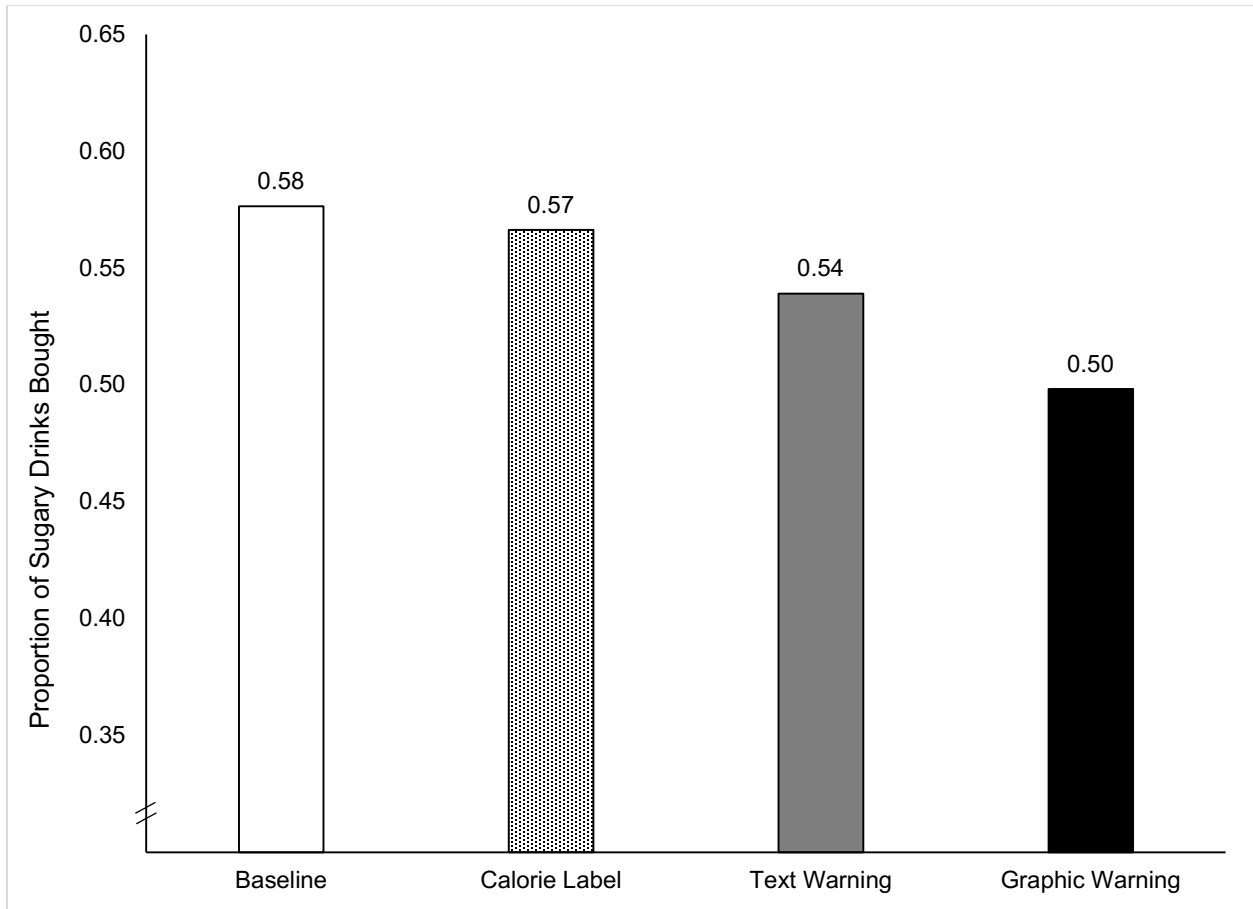


**Appendix Figure D.4** Proportion of bottled drinks purchased per day that were sugary drinks, by condition, in Study 1.



**Appendix Figure D.5** Example of stimulus for Study 2, experimental condition.





**Appendix Figure D.6.** Sugary drink fountain purchases by condition in Study 1.

The proportion of all fountain drinks purchased that were sugary drinks. Fisher's exact tests were used to assess statistical significance, where the unit of observation is a proxy for total drinks purchased: total ounces divided by the average drink size, in ounces. The graphic warning label period resulted in a statistically significant drop relative to baseline ( $p = .01$ ) and the calorie warning label ( $p = .02$ ), but not the text warning label ( $p = .20$ ). No other comparisons are statistically significant (calorie label to baseline:  $p = .76$ ; text warning label to baseline:  $p = .23$ , calorie label to text warning label:  $p = .38$ )



Appendix Table D.1

*Effect of interventions on daily proportion of sugary drinks purchased, unadjusted and controlling for seasonality (Study 1).*

|                 | Model 1             | Model 2             | Model 3             |
|-----------------|---------------------|---------------------|---------------------|
| Calorie Label   | -0.001<br>(0.010)   | 0.002<br>(0.009)    | -0.007<br>(0.012)   |
| Text Warning    | -0.001<br>(0.010)   | -0.010<br>(0.012)   | -0.006<br>(0.021)   |
| Graphic Warning | -0.034**<br>(0.010) | -0.059**<br>(0.023) | -0.063**<br>(0.022) |
| Calendar Week   | ---                 | 1.265<br>(0.927)    | 0.811<br>(1.042)    |
| Heat Index      | ---                 | --                  | -0.001<br>(0.001)   |
| Constant        | 0.219***<br>(0.01)  | -0.041<br>(0.007)   | 0.112<br>(0.248)    |
| Observations    | 56                  | 56                  | 56                  |
| Adj. R-squared  | 0.21                | 0.22                | 0.22                |

*Note.* Each column presents a linear regression estimating the daily proportion of sugary drinks purchased out of all bottled drinks purchased. Robust standard errors are in parentheses. Model 1 is unadjusted. Model 2 controls for calendar week effects. Model 3 further controls for daily heat index. \*\*  $p < .01$ , \*\*\*  $p < .001$

Appendix Table D.2

*Effect of interventions on daily unit sugary drink purchases (number of bottles), unadjusted and controlling for seasonality (Study 1).*

|                        | Model 1            | Model 2             | Model 3             | Model 4             |
|------------------------|--------------------|---------------------|---------------------|---------------------|
| Calorie Label          | 5.64<br>(14.22)    | 5.64<br>(4.63)      | 0.79<br>(5.82)      | 0.58<br>(6.85)      |
| Text Warning           | -4.71<br>(13.23)   | -4.71<br>(5.25)     | -14.52<br>(8.16)    | -14.82<br>(9.61)    |
| Graphic Warning        | -12.36<br>(13.77)  | -12.36†<br>(7.01)   | -19.45*<br>(9.39)   | -19.77†<br>(10.96)  |
| Holiday or Weekend Day | --                 | -69.70***<br>(3.45) | -69.70***<br>(3.48) | -69.75***<br>(2.88) |
| Calendar Week          | --                 | --                  | 0.13†<br>(0.07)     | 0.13†<br>(0.075)    |
| Heat Index             | --                 | --                  | --                  | -0.01<br>(0.17)     |
| Constant               | 77.64***<br>(9.87) | 97.55***<br>(4.33)  | 32.92**<br>(36.70)  | 34.30*<br>(36.64)   |
| Observations           | 56                 | 56                  | 56                  | 56                  |
| Adj. R-squared         | -0.02              | 0.81                | 0.82                | 0.82                |

*Note.* Each column presents a linear regression estimating the units of sugary drinks purchased each day. Robust standard errors are in parentheses. Model 1 is unadjusted. Model 2 controls for whether it is a weekday versus holiday or weekend day. Model 3 adds a control for calendar week effects. Model 4 further controls for daily heat index.

†  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ , \*\*\*  $p < .001$