Field Experiments in Behavioral and Public Economics

Field Experiments in Behavioral and Public Economics

A dissertation presented

by

Syon Pandya Bhanot

to

The Committee on Higher Degrees in Public Policy

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Public Policy

Harvard University

Cambridge, Massachusetts

April 2015

Dissertation Advisors:
Professor Richard Zeckhauser
Professor Brigitte Madrian

Author:
Syon Pandya Bhanot

Field Experiments in Behavioral and Public Economics

Abstract

The three essays in this dissertation present field experiments exploring phenomena in behavioral and public economics in real-world settings.

The first essay outlines a field experiment that uses mailers with peer rank information to motivate water conservation. The essay contributes some of the first pieces of evidence on how comparisons with specific peers might influence behavior. The main finding is that while competitive framing of peer information has positive impacts on efficient homes, it has simultaneous negative impacts on inefficient homes, which are larger in magnitude. In particular, the essay finds that households who rank last in a displayed peer comparison are demotivated by their poor performance, and increase their water use relative to controls.

The second essay studies the impact of signing an explicit promise statement at loan initiation on ensuing loan repayment behavior. The essay provides one of the first field tests of a phenomenon observed in laboratory studies, namely that making a promise can change people's ensuing behavior. Interestingly, the essay does not find support for this claim, and shows the potential difficulty in generalizing laboratory results to real-world settings.

The third essay focuses on decision making about risk. Specifically, it presents two field studies that use quasi-random, real-world events to explore how emotions influence risk decisions. These studies are among the first field tests of the relationship between emotion and risk preferences. The essay offers mixed results, finding that negative emotions seem to increase risk aversion only when the emotions derive from events linked to individual self-responsibility.

Contents

Acknowledgments

I am deeply indebted to my friends, family, academic advisors, and colleagues for supporting me as I completed this work.

On the academic side, I am grateful for the tireless support of Richard Zeckhauser and Brigitte Madrian, my primary advisors. Richard and Brigitte supported me in exploring the topics I cared about, and offered guidance at every step (and patience at every misstep). The same could be said for the other academics who assisted me at various stages in the process, including: Alberto Abadie, Dan Ariely, Amitabh Chandra, Gary Charness, David Johnson, Steven Levitt, Dan Levy, John List, Greg Mankiw, Janina Matuszeski, Tim McCarthy, Michael Norton, and Monica Singhal.

I have also been fortunate enough to collaborate with some extremely smart people, including: Vivien Caetano, Chaning Jang, Gordon Kraft-Todd, Julia Lee, Benjamin Lockwood, David Rand, Matthew Ranson, Jeremy Shapiro, and Erez Yoeli. I am looking forward to continuing these collaborations in the months and years to come.

My friends have tolerated me for years—and long may they continue to do so. A special thanks to David Bartels, Devan Darby, Kant Desai, Jake Gordon, Chloe Green, Tali Malott, Ernest Mitchell, Safiyy Momen, Jamal Motlagh, Sahil Raina, Rob Rich, David Sheng, and Nitin Walia for keeping me sane. And to Susanne Schwarz—your patience, advice, and affection contributed to every word in these pages. I was lucky to have you by my side, and I hope you'll always be there.

Last and not least, my parents and siblings made this happen. Gyan Bhanot, Rashmi Bhanot, Kavita Bhanot, and Meru Bhanot: this is for you. Oh, and thanks to Humphrey as well, for being the world's best canine alarm clock.

To my parents, Gyan Bhanot and Rashmi Bhanot.

Introduction

> "There are some situations which bear so hard upon human nature, that the greatest degree of self-government, which can belong to so imperfect a creature as man, is not able to stifle, altogether, the voice of human weakness, or reduce the violence of the passions to that pitch of moderation, in which the impartial spectator can entirely enter into them. Though in those cases, therefore, the behaviour of the sufferer fall short of the most perfect propriety, it may still deserve some applause, and even in a certain sense, may be denominated virtuous."
>
> - Adam Smith, *The Theory of Moral Sentiments*, 1759[1]

Behavioral economics is the study of the psychology of economic decision making. It is a field with deep roots, but growing branches.[2] While many regard the surge in behavioral work in the last half century, particularly that of Amos Tversky and Daniel Kahneman, as the genesis of the field, the origins of behavioral economics can actually be traced back hundreds of years. Adam Smith's *The Theory of Moral Sentiments* (published in 1759 and sometimes overlooked in favor of his iconic 1776 text, *The Wealth of Nations*) spelled out many of the behavioral concepts academics study to this day (Ashraf *et al.*, 2005). Thus, while Smith is often credited with the view that human beings can serve society most effectively by acting as rational, self-interest-maximizing agents, guided by the "invisible hand" of the market, his full body of academic work paints a more nuanced picture. Human beings are complex, and our decisions are often driven by "human weakness" and "passions" in ways

[1]Smith (1759)

[2]For excellent overviews of the history of behavioral economics and decision-making research, see Camerer and Loewenstein (2004) and Fox (2015).

that conflict with pure reason.

In the two centuries following the publication of *The Theory of Moral Sentiments*, many of its insights found a more welcoming home in psychology than in economics. Pioneering academics in 19th century psychology, such as William James, showed a greater willingness than economists to explore the often unobservable phenomena of "feelings, desires, cognitions, reasonings, decisions, and the like."[3] Meanwhile, the discipline of economics, driven largely by the neoclassical revolution of the late 19th and early 20th century, sought to turn economics into a natural science, where human behavior was described using formal models with mathematical structures. This led economists to embrace the model of the rational decision maker, guided by Benthamite utility maximization and self-interest. As the early neoclassical economist W. Stanley Jevons wrote in his 1871 text, *The Theory of Political Economy*, "Pleasure and pain are undoubtedly the ultimate objects of the Calculus of Economics. To satisfy our wants to the utmost with the least effort... in other words, to maximise pleasure, is the problem of Economics."[4] The increasing formalism in economics also fostered a general distaste among 20th century economists for the discipline of psychology. As George Loewenstein and Colin Camerer write, "economists thought [psychology] proved too unsteady a foundation for economics... [which] led to a movement to expunge the psychology from economics."[5]

Following World War II and into the latter half of the 20th century, things slowly began to change. A multi-disciplinary approach to studying decision making and human behavior began to emerge, driven by the work of John von Neumann, Herbert Simon, Ward Edwards, and Howard Raiffa, among others. These scholars built entire fields of study around decision analysis and behavioral decision making, and paved the way for the reintroduction of psychology into the economics discipline. The most well-known of these scholars are the psychologists Amos Tversky and Daniel Kahneman, who used experiments to study

[3]James (1890, p. 1)

[4]Jevons (1871, Chapter III, Section i)

[5]Camerer and Loewenstein (2004, p. 5)

heuristics and biases, violations of expected utility, and other anomalies inconsistent with the rational model of economic behavior. Their collaborative work, which began in the 1970s and continued for nearly 30 years, transformed the academic landscape around decision making, and earned Kahneman a Nobel Prize in economics in 2002.[6] This research, in turn, inspired the first generation of "behavioral economists," who to this day explore how humans make decisions and refine economic models to incorporate their findings.

The integration of psychological insights into economics has not been smooth, however, with bickering between behavioral and neoclassical economists a recurring sight. Many of these ongoing debates miss the mark, pitting the two perspectives against one another rather than portraying them as two useful models for thinking about decision making in different contexts. Much of this is, as Richard Zeckhauser describes, a deeply-entrenched "turf war."[7] However, some of this tension stems from behavioral economics' traditional reliance on controlled laboratory experiments, which many critics see as limited in their generalizability. As economist John Cochrane says, "These (behavioral economics) experiments are very interesting... The next question is, to what extent does what we find in the lab translate into understanding how people behave in the real world? And... does this explain market-wide phenomena?"[8]

This is a very good question—do the insights from behavioral economics actually explain real-world behavior? In recent years, behavioral economists have increasingly shed their lab coats for work boots to answer this question, venturing into the field to test behavioral concepts using randomized experiments.[9] The insights from these endeavors have been profound, not just for behavioral economics, but for the field of economics more broadly. One of the pioneers of this field experimental approach is John List, whose wide array of experimental work includes studies that have changed the way we think about how

[6] Tversky did not receive the same honor, because he died in 1996. The Nobel Prize is not awarded posthumously.

[7] Zeckhauser (1986, p. S436)

[8] Cochrane (2010)

[9] For a comprehensive summary of this work, see DellaVigna (2009).

behavioral concepts play out in the real world. For example, his work suggests that the endowment effect may be tempered by market experience in real-world settings (List, 2003), and that charitable giving in door-to-door fundraisers is largely a function of social pressure (DellaVigna *et al.*, 2012). Additionally, a growing body of development economics research has centered around using field experiments, often inspired by behavioral economics, to test various policy interventions in the developing world.

The use of field experiments in behavioral science is especially important because of the field's increasing prominence in public policy circles. Indeed, the last decade has seen a meteoric rise in the concentration of behavioral researchers working in governments and non-governmental agencies worldwide. For example, in 2009, U.S. President Barack Obama appointed Cass Sunstein, a Harvard Law School professor and co-author of the influential book "Nudge," which advocates for the use of behavioral science to improve policy and regulation, as the Administrator of the White House Office of Information and Regulatory Affairs. Meanwhile, in the U.K., Prime Minister David Cameron created the "Behavioral Insights Team"—also known as the "Nudge Unit"—as part of his Cabinet Office in 2010.[10] This group is tasked with using experimental methods to test how behavioral economics and social psychology can be used to improve public policy. The U.S. government followed suit in 2013, when Maya Shankar was selected to lead the Social and Behavioral Sciences Team as part of the White House Office of Science and Technology Policy. Similar efforts to incorporate behavioral science into government practice are underway in Canada, Denmark, Australia, Germany, and France, among others.[11]

While the growing influence of behavioral science in public policy seems likely to improve policymaking, it is important that researchers involved with these policy efforts do not overreach in making claims about the efficacy of behaviorally-motivated interventions. Many behavioral concepts have yet to be tested in real-world settings. Furthermore, despite

---

[10]The Behavioral Insights Team has since been partly privatized, and is now co-owned by its employees, the U.K. Government, and Nesta, an innovation charity in the U.K.

[11]Subramanian (2013)

recent growth in the area, the existing body of field experimental evidence in public economics, the subfield of economics concerned with the role of the government in the economy, is relatively small. If behavioral researchers are going to have a lasting impact on public policy, they will need to further expand the body of field research showing that behavioral concepts actually work in policy-relevant settings, including those related to the major questions in public economics.

Dissertation Outline

This dissertation presents three papers that contribute to this important area of behavioral and public economics research. While the papers I present touch on different behavioral concepts, all three explore areas with limited real-world evidence in existing literature.

The first paper, "Rank and Response: A Field Experiment on Peer Information and Water Use Behavior," studies how information about the people around you can influence your cooperative behavior. It does so using a natural field experiment, testing how and why mailers with social information encourage water conservation in Castro Valley, California. While there has been significant work by social scientists on social information in the last half century, there is less work on what underlying motivators might explain observed responses, or on how comparisons with specific peers might influence behavior. This paper provides some evidence in these areas, finding that while competitive framing of peer information has positive impacts on efficient homes, it has simultaneous negative impacts on inefficient homes, which are larger in magnitude. This study can be thought of as broadly exploring how learning about the behavior of other people can influence us to cooperate in real-world public goods situations—scenarios where everyone needs to "do their part," but everyone also has an incentive to free-ride on the pro-social behavior of others.

The second paper, "Cheap Promises: Evidence from Loan Repayment Pledges in an Online Experiment," studies the impact of signing an explicit promise statement at loan initiation on ensuing loan repayment behavior. Again, this study uses a natural field experiment, conducted with a partner firm in the business of online lending. The paper provides one of the first field tests of a phenomenon observed in laboratory studies, namely that making a promise—even if it is not enforceable—can change people's ensuing behavior. Interestingly, this study does not find support for this claim, and shows the potential difficulty in generalizing laboratory results to real-world settings. The results here are important not only because they test existing theory, but also because they provide experimental evidence about how behavioral interventions can (or, sometimes, cannot) influence financial decision making. This is of growing policy significance, with organizations like the U.S. Consumer Financial Protection Bureau and the U.K. Financial Conduct Authority increasingly using behavioral insights to improve public policy around consumer financial products.

The third paper, "Does Negative Emotion Increase Risk Aversion?: Evidence from Exam Grades and NFL Games" (co-authored with Julia Lee and Matthew Ranson), focuses on the psychological origins of decision making about risk. Whereas the two previous papers rely on a randomized experimental methodology, with subjects unaware of the experimental manipulation, this paper studies recruited subjects and uses quasi-random, real-world events to explore the link between emotion and risk preferences. It therefore presents an alternative methodology for bringing behavioral results from the laboratory to the real world for testing. That is, by using naturally-occurring events as the experimental manipulation, along with appropriate panel instrumental variable techniques, we can better study "genuine" emotion and its effects on behavior. The two field studies presented in the paper offer mixed results, finding that negative emotions seem to increase risk aversion only when the emotions derive from events linked to individual self-responsibility. These results represent some of the first pieces of field evidence on the link between emotion and risk taking.

Understanding Mixed Results

Notably, the three papers I present in this dissertation show two stark realities about experimental work, particularly in field contexts.

First, it is often the case that the results of experiments are mixed, contrary to what was expected ex-ante, or non-existent (in the form of "null results"). This presents a challenge to the experimentalist. What should be done in the face of mixed or null results? From the perspective of academic discourse, generally speaking, such results should be shared with the academic community, their publication should be celebrated as a triumph for honest research, and any inconsistencies in the results should drive further academic inquiry.

However, more often than not, this is not what happens. In Franco *et al.* (2014), researchers analyzed the publication rates of experiments conducted using the Time-sharing Experiments in the Social Sciences (TESS) program, which is sponsored by the National Science Foundation. They find strong evidence of publication bias based on the statistical strength of experimental results. Specifically, Franco *et al.* (2014) report that statistically-significant results are 40 percentage points more likely to be published than null results. Just as troubling, the researchers find that 65 percent of "null result" papers are never even written up by researchers. The self-censoring of results is likely driven in part by the incentives in academia, summarized by the colloquialism: "publish or perish." However, independent of the reasons underlying the behavior, the failure to write papers reporting null or mixed results deprives the academic community of valuable signals about the existence (or lack thereof) of critical effects or relationships between variables. Simply put, if such publication bias is pervasive, then even the keenest consumer of social science research cannot know whether anything they read is true.

What can be done about this problem, as it pertains to experimental work in the social sciences? The obvious solution is perhaps one of the hardest to implement. That is, academic journals should soften their often hard-line stance on statistical significance, or at least show a greater willingness to publish mixed or null results in cases where an experiment is deemed sufficiently well-run or academically important. To be sure, some journals are more progressive in their thinking about this issue than others (with the JASNH, the Journal of Articles in Support of the Null Hypothesis, a rather extreme example). Furthermore, there are certainly cases of established social science journals publishing null or mixed results (examples include Banerjee *et al.* (2008) and Dustmann and Schonberg (2012)).[12] However,

[12]An alternative to publishing null results would be to have a website that allows researchers to post their

by and large, it seems fair to state that most journals suffer from such publication bias, and there is no obvious solution to this urgent problem forthcoming.

Second, social scientists must be cautious about designing experiments whose results can be ex-post justified, regardless of the direction of the observed effect. Such designs can cause bias of another kind, from researchers altering their written predictions to match the results—a phenomenon that can be dubbed the "promiscuous prediction problem."[13] In such cases, the concern can be captured by a simple question: If any effect from a given experiment can be justified as having a deep scientific meaning, does the experiment have any value ex-ante?

It is here that the importance of the academic process for experimental research becomes extremely important. First and foremost, experiments should, whenever possible, be explicitly linked to existing theory and hypotheses ex-ante, ideally through mechanisms like online registries for submitting hypotheses and pre-analysis plans. One such example is the American Economic Association's registry for Randomized Control Trials, established in 2012. This registry provides a platform for experimentalists to submit analysis plans in advance of receiving data, to ensure that they are not tempted to manufacture results. Of course, using a pre-analysis plan posted on such a registry does not mean a researcher cannot modify their analysis as they go. Indeed, many phenomena are more complex than a researcher may imagine prior to an experiment, with conflicting behavioral forces or unexpected heterogeneous effects possible. Nevertheless, such registries provide some discipline, and would represent a major improvement over a "black box" approach that lends itself to researcher manipulation.

A second viable mechanism for addressing this issue is simple academic humility. By couching results in the broader context of existing literature and not overreaching in asserting the definitiveness of a single paper's contribution, academic researchers can help

---

null results in a searchable database, along with information about any related professional presentations or working papers.

[13]Full credit for this apt term should go to Harvard's Richard Zeckhauser.

address the problem in-house.

These are especially salient challenges for field experimentalists. One could argue that the "messy" nature of field experiments, in terms of linkages to theory and the challenges of external validity, makes such studies more prone to null, mixed, or unclear results. Furthermore, since field experiments are not commonplace in many subfields of social science, it is often the case that new field experimental findings represent "the first" evidence from the real world on a given topic or concept. Therefore, field experimentalists usually cannot point to a large body of "existing field evidence" on similar interventions to corroborate their findings, relying instead on theoretical, empirical, or lab work (which are quite different methodologically). How should field researchers move forward, then? One solution may be for field experimentalists to come back to the lab, or to existing data sets, more often, to validate theories or insights obtained in field work. By showing that a specific unexpected finding in a field experiment can be replicated in a controlled laboratory setting, field researchers can significantly enhance the validity of their results and contribute much more to existing literature. A second solution is similar to the one presented above for dealing with mixed or null results; that is, field experimentalists should display caution when discussing their findings, particularly when they know they may be basing their conclusions on "promiscuous predictions."

With these challenges in mind, the essays in this dissertation seek to find the right balance between conclusiveness and caution, as the experimental data and analytical methods warrant. In finding that balance, I hope to contribute experimental evidence and insights from real-world settings, in order to improve our understanding of how "nudges" and motivators not traditionally studied by economists (like emotions and social norms adherence) can influence behavior. Importantly, my goal is not to prove that behavioral interventions are always impactful—indeed, some of my findings suggest that such interventions can fail to influence behavior in the ways predicted by laboratory studies and existing behavioral theories. Instead, I aim to develop and test ideas that simultaneously inform real-world policymaking and refine formal models in both behavioral and public economics.

Chapter 1

Rank and Response: A Field Experiment on Peer Information and Water Use Behavior

1.1 Introduction

Research in psychology and behavioral economics has consistently demonstrated that our social surroundings affect our behavior and that we are influenced by how we compare to our peers (Schultz *et al.*, 2007; Griskevicius *et al.*, 2008; Fehr and Gintis, 2007; Beshears *et al.*, 2015). Traditionally, economists studying decision making have focused less on these social motivators and more on financial ones. However, financial incentives are unable to change behavior in some contexts, either because pricing is not salient or because behavioral elasticities are low (Ferraro *et al.*, 2011; Allcott, 2011; Olmstead and Stavins, 2007). In these cases, it may be beneficial to stimulate behavior change with social motivators, like the desire to attain a high rank relative to our peers (Tran and Zeckhauser, 2012). In this paper, we present an experiment that tests the effect of social rank on behavioral response, and explores how the framing of peer information can influence this response.

Experimental researchers have studied how social information can alter behavior in a variety of contexts, including energy conservation, voting, and savings (Allcott, 2011; Gerber *et al.*, 2008; Kast *et al.*, 2014; Beshears *et al.*, 2015).[1] Most interventions have provided individuals with information on the average performance of a broader social group, with mixed results. Allcott (2011), for example, finds that showing individuals how their energy use compares to the mean of both their most efficient neighbors and all of their neighbors reduces electricity consumption in the average household by over 2%. However, other research suggests that sharing peer information can lead to socially undesirable behavior. Beshears *et al.* (2015) find that the provision of peer information about savings for retirement can reduce savings rates by demotivating low-performers. John and Norton (2013) document a related phenomenon in the context of workplace exercise "walkstations." They find that people tend to converge to the bottom performer, exercising less at walkstations when given information about the low rates of use by others. Another set of studies on the use of social information to influence alcohol abuse on college campuses find no discernable effect of such messaging on overall outcomes (Wechsler *et al.*, 2003; Clapp *et al.*, 2003; Granfield, 2005).

One limitation of existing work is that it does not disentangle the various motivating and demotivating elements of social information, and does not isolate those that are central to behavioral responses. When social information works, is it because the information primes our competitive drive or because it stimulates our cooperative spirit? The literature also does not say a great deal about peer comparisons, explicit rank information, or heterogeneities in their motivational effects. How do ranked comparisons to specific people who are "like us" motivate us differently than aggregate social comparisons? We hope to offer insights into these questions.

In this paper, we outline a natural, randomized field experiment that tests how peer

[1] Beshears *et al.* (2015) identifies a number of additional studies in this space, noting that, "providing information about peers moves behavior towards the peer norm in domains such as entree selections in a restaurant, contributions of movie ratings to an online community, small charitable donations, music downloads, towel re-use in hotels, taking petrified wood from a national park, and stated intentions to vote (Cai, Chen, and Fang, 2009; Chen et al., forthcoming; Frey and Meier, 2004; Salganik, Dodds, and Watts, 2006; Goldstein, Cialdini, and Griskevicius, 2008; Cialdini et al., 2006; Gerber and Rogers, 2009)." (Beshears *et al.*, 2015, p. 1)

information influences behavior.[2] The experiment was conducted with a partner firm in California, which works with local utilities to reduce water use at the household level through household mailers and other outreach campaigns. In the experiment, we used mailers with different forms of peer information and social rank messaging to motivate reductions in water use. Through the experimental design, we can address existing theories about how peer information, peer comparison, social rank, and the framing of behavioral messaging can influence behavior. The goal of this study is to provide insights that contribute to an improved, coherent theory about social information and its potential for heterogeneous effects.

Our results suggest that while social information can reduce water use, peer ranking and framing can have detrimental impacts on behavior. Specifically, we find overall water use reductions in the range of 13-17 gallons per day in response to a four-piece mailer campaign containing different peer information and social rank content. However, we find evidence of heterogeneity in treatment effects from rank information. In particular, households that were low water users prior to the experiment showed a "boomerang effect" (i.e., an increase in water use) from rank information, except when a competitive frame was included. This result is consistent with Garcia *et al.* (2006), who posit that receiving high rankings can spur competitiveness in a way that makes people less likely to "boomerang." However, the competitive frame had detrimental effects on the behavior of households that were high water users prior to the experiment, demotivating them and increasing their water use on average. Further analysis of rankings suggests the existence of a "last place effect," whereby competitively-framed rank information led to an increase in water use by the worst

[2]We refer to four forms of behavioral messaging in the paper: social information, peer information, peer comparison, and social rank. We define and distinguish between them as follows. Social information is the broadest category, referring to any messaging containing information about the behavior of others. Peer information is a subset of social information, referring to messaging that conveys information about a given individual's behavior and information about their peers (people "like them"). This term encompasses information conveyed either at the aggregate level ("here's how your peers performed on average") or in a more detailed manner ("here's how you performed relative to a similar household"). Peer comparison is a subset of peer information, referring to the display of the specific outcomes of an individual and their peers, provided explicitly at the individual level. Finally, social rank refers to messaging that informs individuals of their hierarchical position among peers.

performer in the peer group—a movement away from the social norm. We believe this stems from the potentially demotivating power of peer information, in line with the results on peer information and savings in Beshears *et al.* (2015).

The paper proceeds as follows. Section 1.2 provides background on existing work, related theories, and the policy context for this intervention. Section 1.3 provides details on the experiment. Section 1.4 presents the empirical methods used to analyze the data from the experiment. Section 1.5 provides results. Section 1.6 provides a brief discussion and concludes.

1.2 Background

Water conservation provides an important context to test the effects of social information, since individual water use behavior is both important to change and difficult to influence. Water leaks provide a useful illustrative example. In a 1999 study, the American Water Works Association Research Foundation found that nearly 14% of household water use comes from leaks.[3] The leak problem is diffuse—the EPA estimates that roughly 10% of homes have leaks that waste 90 or more gallons of water each day.[4] Fixing a leak can significantly reduce a household's water use. Yet few households seem to be fixing their leaks.

The absence of such efforts can be partly explained by the human tendency to only process salient information. Water leaks are generally invisible, requiring professional assistance to find and fix.[5] This salience problem is further compounded by the relatively low price of water—the average family in the United States spends only 0.5% of household income on water and sewage bills.[6] Given the lack of salience and the low price, it is not

[3] Mayer and DeOreo (1999)

[4] WaterSense (2015)

[5] American Water has a Leak Detection Kit online that outlines common indoor leaks in detail and explains how to find them. See: http://www.amwater.com/files/AMER0231_LeakDetectionKitWeb_Layoutopt.pdf.

[6] EPA (2009)

surprising that the price elasticity of water is low. Olmstead and Stavins (2007) estimate a water price elasticity of -0.33 and point to an earlier analysis by Espey *et al.* (1997), which placed 90% of all estimates between 0 and -0.75. More recent estimates in California in Lee and Tanverakul (2015) find elasticities in the -0.2 to -0.5 range, suggesting that water prices would have to increase significantly in our experiment's target areas to have a major impact on water conservation decisions.[7]

Households are also unlikely to change their water use without knowledge of what constitutes "good" and "bad" water consumption behavior in their community. Generally, households do not receive such information. However, social information interventions may offer a solution, by providing a reference point for individuals to evaluate their water use.[8] There is evidence that providing social information in this way can be a simple and low-cost means of changing behavior (Jessoe and Rapson, 2014; Dickerson *et al.*, 1992). Given this, we designed our experiment to explore the influence of social rank and peer information in the context of a water conservation mailer program.

1.2.1 Existing Theories on Rank and Response

A number of important theories from social science literature might explain how social rank and peer information affect behavior—with very different predictions. A brief discussion of these theories and their predictions is presented here.

Social Norms Theory

Social norms theory predicts that peer information motivates behavior change because it provides a social standard to follow. Most notably, the theory of social comparison processes

---

[7]In Allcott (2011), providing social norms and information decreased energy use by roughly the same amount as a 11-20% increase in price.

[8]Another potential danger of not providing households with such information is heuristic decision making, which has been studied in the energy conservation literature (Gillingham *et al.*, 2009). When households do not have enough information about their energy use, they may rely on heuristics to determine their energy consumption, which a number of experiments found led to miscalculations of use and overconsumption (Kempton *et al.*, 1992; Kempton and Montgomery, 1982).

presented in Festinger (1954) suggests that social comparison occurs when objective, non-social standards are unavailable. This could lead individuals to evaluate their opinions and abilities by comparing themselves to others—and to take action to reduce any found discrepancies. Furthermore, Festinger argues that individuals are most likely to compare themselves to people who are similar to them.[9] Schultz *et al.* (2007) similarly state that social information can send the message that "being deviant is being above or below the norm."[10] Social norms theory then implies that providing individuals with rank information would cause their outcomes to compress towards the displayed social norm.

In recent years, there have been an increasing number of experimental tests of these theories. For example, Schultz *et al.* (2007) conducted a field study with several hundred households in San Marcos, California, using door hangers with aggregate-level social information on energy use to motivate energy reduction. They find that social information caused high energy use households to decrease their energy use, but encouraged low energy use households to increase energy use. On the one hand, this implies a desirable response to social information from low-performing individuals. However, it also predicts a detrimental response from high-performers, referred to as the "boomerang effect."

The boomerang effect hypothesis has its roots in the psychology of motivation. Proponents of the effect suggest that a favorable social comparison provides a license for high performers to behave worse (Clee and Wicklund, 1980). For example, telling individuals that most people in their workplace do not put in overtime sends the message that putting in overtime is unnecessary. The boomerang effect has been documented in some interventions focused on social rank, but there is a relative shortage of experimental evidence on the effect (Fischer, 2008). In this experiment, we explore how framing and context might influence the boomerang effect.

It is important to note that any observed compression towards the mean could be

[9]For a discussion on the specific variables relevant for comparison (e.g. expertise, similarity and previous agreement), see Suls *et al.* (2002).

[10]Schultz *et al.* (2007, p. 430)

attributed to mean reversion, which can happen in any experimental setting in the absence of messaging with social norms (Kahneman, 2011). We must be careful when assessing interventions using social rank information to ensure that we do not confuse mean reversion with responses to social norms, particularly in the short run. Using a randomized experiment that varies whether subjects are exposed to specific social norms information helps us distinguish between these effects.

Motivation Effects

Academic literature on motivation and self-efficacy suggests another possibility: that individual outcomes will widen away from the mean, as those who rank well among their peers will work harder to improve and those who rank poorly will "give up."[11] There is a rich body of research underpinning this prediction in the social sciences. Research with children has shown that our beliefs about our abilities are influenced by what others tell us.[12] If we feel—or are told—that we are good at an activity, we are more likely to engage in it, whereas we avoid activities for which we feel ill-equipped (Bandura, 1977).[13] Research in exercise and sport performance has shown that verbally reciting instruction messages that convey positive beliefs improves ensuing performance outcomes (Shelton and Mahoney, 1978). This suggests that individuals with positive beliefs about their ability may set higher goals for themselves and try harder to achieve them. This is commonly attributed to the view that "high efficacy" people view difficult or new tasks as challenges rather than threats (Bandura, 1994; Yim and Graham, 2007).

Individuals with low-efficacy (those who receive low-rankings), on the other hand, may quit once they learn of their poor rank (Hagger *et al.*, 2002). Beshears *et al.* (2015) found that low-savings individuals were discouraged by information about peers' savings rates,

[11] For a summary of literature in this area, see Pajares (1997).

[12] NASP (2010)

[13] Education research has notably used this theory to suggest that teacher efficacy, or a teacher's belief in his or her ability to bring out the best in students, has powerful effects on student achievement, student motivation, and teacher behavior (Tschannen-Moran and Hoy, 2001).

Table 1.1: *Rank Response Predictions and Supporting Theories*

Predicted Response to Rank	Supporting Literature/Theories
Top Performers Improve	Self-Efficacy (Bandura, 1977, 1994) and Rank Competition (Garcia *et al.*, 2006)
Top Performers Worsen	Social Comparison (Festinger, 1954) and Social Norms (Clee and Wicklund, 1980; Schultz *et al.*, 2007)
Bottom Performers Improve	Social Comparison (Festinger, 1954) and Social Norms (Clee and Wicklund, 1980; Schultz *et al.*, 2007)
Bottom Performers Worsen	Self-Efficacy (Bandura, 1977, 1994), Demotivation (Hagger *et al.*, 2002; Beshears *et al.*, 2015; Polivy and Herman, 1985), "Loss Demotivation" (based on Kahneman and Tversky (1979))

which the authors attributed to the discouraging effects of upward social comparison. This response, now commonly referred to as the "what the hell effect," was also identified in dieters by Polivy and Herman (1985). The authors found that once a dieter exceeds their caloric intake goal for a single day, they proceed to eat significantly more calories than the goal. In short, they perceive their performance as a failure and respond by "binging."

The prediction of demotivated low-performers also finds support in loss aversion, which posits that a given loss affects the psyche more than an equivalent gain (Kahneman and Tversky, 1979). In this context, a poor performance relative to the social norm may be perceived as a "loss," and the worst-performing individuals may feel demotivated by the impossibility of catching up. The high emotional weight of losses may therefore bring "loss demotivation." Taken together, the what-the-hell effect, self-efficacy theories, and loss demotivation suggest that providing information on social rank may cause the spread of outcomes to widen, with those at the top inspired to try harder and those at the bottom giving up.

1.2.2 Social Messaging Frames

Literature in psychology, behavioral economics, and marketing has reliably found that the framing of information can alter behavior (Winter, 2008). While existing work has not explored framing in the context of social rank specifically, there are related results that provide testable predictions. For example, Schultz *et al.* (2007) argue that the "boomerang effect" can be counteracted if an injunctive message, which conveys social approval or disapproval, is included. The authors found that low energy users receiving an injunctive message maintained their low use rates, while those who did not suffered a boomerang effect.[14] Similarly, Allcott and Mullainathan (2010) attribute the lack of a boomerang effect in their experiment to the use of injunctive messaging.

In this experiment, we use both competitive and cooperative frames to explore the nuances of how social information influences behavior, building on past work. For example, in an experiment exploring the effect of negative and positive frames of cooperative messaging on behavior, Cialdini *et al.* (2006) found that positive framing of cooperative messaging encouraged people to cooperate. Conversely, negatively-framed cooperative messaging provided people with a justification for their own bad behavior. Other research suggests that cooperative frames can backfire when there is no widespread rule adherence, as individuals might use the cooperative setting to free ride (Olson, 1965).

Framing social comparison as a competition, through status rewards such as medals, can also be powerful. Online and social media giants like Foursquare, Overstock, Yelp, and Wikipedia all use non-financial status rewards to motivate users. These rewards prime our competitive desire to obtain a higher social rank, and can serve as a form of personal affirmation that increases self-efficacy (Antin and Churchill, 2011). Furthermore, Garcia *et al.* (2006) argue that ranking may itself drive competitiveness, finding that individuals are most competitive when they or their competitors are highly ranked. They also argue that the degree of competition between rivals "depends on their proximity to a meaningful

[14] In Schultz *et al.* (2007), the injunctive message used was a "happy face" if subjects used less energy than the social average and a "sad face" if subjects used more energy than the social average.

standard."[15] Therefore, amongst top performers, rankings and competitive framing may mutually reinforce in a way that motivates positive behavior change.

Meanwhile, we might expect low-performing individuals to change their behavior to avoid "being last," particularly if they are close to the margin, due to "last place aversion" (Kuziemko *et al.*, 2014). However, research on goal-setting and attainment suggests that the use of competition and rank can also be demotivating. For example, both Beenen *et al.* (2004) and Harding and Hsiaw (2014) suggest that individuals may do worse if they feel that their target goals are unachievable. Similarly, Little (2012) finds that competitive frames can demotivate individuals if they reinforce patterns of failure.

Overall, there is little consensus on the relative value of cooperative versus competitive framing (Little, 2012; Qin *et al.*, 1995; Kohn, 1996; Julian and Perry, 1967). While this experiment will not resolve these debates, it contributes experimental evidence in a specific context that can inform further research on the use of cooperative and competitive frames to motivate behavior change.

1.2.3 Policy Context

Water shortages and drought are serious and worsening problems across most of California. In January 2014, California Governor Jerry Brown declared a drought state of emergency in the state, calling for "all Californians to conserve water in every way possible."[16] U.S. President Barack Obama visited the state shortly thereafter; during a February 2014 visit to Los Banos, CA, Obama stated, "everybody, from farmers to industry to residential areas [... is] going to have to start rethinking how we approach water for decades to come."[17] The situation continued to worsen, however, with drought levels rising to unprecedented heights in early 2015. Governor Brown responded by taking a historic step in April 2015, mandating a 25% reduction in water use across the state—the first such mandatory reduction in the

[15] Garcia *et al.* (2006, p. 970)

[16] Brown (2014)

[17] Obama (2014)

state's history.[18]

This is not a problem unique to California. At present, roughly one billion people worldwide lack access to safe drinking water. Increasingly, experts warn that the persistent shortage of water resources will have ramifications not only for human health and the environment, but for political stability and national security.[19] Based on the notion that small efforts at reduction at the individual level can have major aggregate impacts, water conservation efforts have focused on the overuse of water at the household level.

Some argue that directly restricting water use or raising prices may be the answer, but these two options have flaws. Government-imposed restrictions on water use may make individuals less enthusiastic about conserving water, undermining long-run behavior change (Lynne *et al.*, 1995; Watson *et al.*, 1999). Furthermore, a report by the Pioneer Institute concludes that such programs can also be expensive, especially considering that empirical evidence regarding their aggregate effects is mixed (Olmstead and Stavins, 2007). Price-based approaches and subsidies are similarly underwhelming. Hunt Allcott argues that they are politically infeasible, difficult to evaluate, and "in practice a large drain on increasingly-limited public funds."[20] The problem of price inelasticity discussed earlier also means that the impact of price-based regulation would be limited. This suggests a possible role for behavioral interventions in water conservation policy.

1.3 Experiment Overview

1.3.1 Motivating Literature for Experiment Design

With this field experiment, we seek to understand the mechanisms of social information effects, and how framing affects behavior change. The basic design features of the experiment draw and build on existing studies on social information, including in the water conservation

[18] Brown (2015)

[19] DNI (2012)

[20] Allcott (2011, p. 1082)

context (Petersen *et al.*, 2007; Davis, 2011; Allcott, 2011; Ferraro and Price, 2013; Allcott and Rogers, 2014). We extend existing work by framing peer information using both cooperative and competitive frames in some of our treatment groups. These frames were chosen to prime different elements of social information, which we argue might influence individuals differently. For example, cooperative frames motivate behavior change by inducing norm compliance, whereas competitive frames motivate behavior change by making an action seem valuable (Fehr and Gintis, 2007).

Additionally, households received multiple mailers in this experiment, which is uncommon in the literature. Most existing experimental work on water conservation has used single mailers sent to households (Ferraro *et al.*, 2011; Ferraro and Miranda, 2013). Furthermore, studies that do use multiple mailers in other contexts tend not to address the effect of additional mailers (Karlan and Zinman, 2009; Bertrand *et al.*, 2010). One exception is Allcott and Rogers (2014), which finds that households responded to repeated mailer treatments over two years, with some decay in effects when mailers ceased. In this paper, we assess both the short- and long-run impacts of the treatment during the experimental period.

1.3.2 Experimental Design

Partners

My research partner for this field experiment was a firm based in California that works directly with public water utilities to promote more efficient water use by California homeowners.[21] The firm sends a personalized mailer, called a Home Water Report (HWR), to households every two months. The HWRs are transmitted either electronically or through traditional mail, and incorporate messages designed to engage customers and reduce water use.[22] Through the utilities, the firm tracks water use and customer engagement over time.

[21]The firm will remain anonymous in this paper.

[22]Approximately 10% of customers receive the Home Water Report by email, with the rest receiving paper mailers.

The public utility partner in the study was a local water provider that serves a subset of homes in the greater San Francisco Bay area. The utility provided the water use data for analysis of the experiment's impact.

Subjects

We conducted the field experiment in Castro Valley, a town of 60,000 residents in Alameda Country, California, roughly 15 miles southeast of Oakland. Subjects in the study were residents of 5,180 single-family households in the C2A pressure zone in Castro Valley, who receive water through the public utility.[23] The panels in Appendix A.1 show the specific location of both Castro Valley and of the C2A pressure zone within Castro Valley specifically. Prior to the start of this experiment, the firm was already working with roughly 4,000 households in the other pressure zones in Castro Valley. This study specifically targeted households in the C2A pressure zone, who were being added to the existing base of Castro Valley customers. This is critical—all subjects in this experiment were receiving Home Water Reports for the first time.

Study Design

The 5,180 households in the experiment were first subdivided into 20 "cohorts" based on two categorical variables: 1) irrigable area; and 2) the number of occupants in the household. There were four possible irrigable area sizes for a household (small, medium, large, and extra large), with irrigable area computed by the firm using real estate data on lot size and home footprint from DataQuick. There were five possible household occupant "buckets" (1, 2, 3, 4, and 5+). Therefore, with four possible "irrigable area" values and five possible "occupant" values, there were 20 possible cohorts for households in the experiment. Table 1.2 outlines the number of households in each of the cohorts.

Every household was individually assigned a random subset of four households in their

[23] A "pressure zone" is a geographical area defined by the public utility based on the area's elevation above sea level.

Table 1.2: *Number of Households by Cohort*

Irrigable Area	1 Occupant	2 Occupants	3 Occupants	4 Occupants	5+ Occupants
Small	350	867	1,381	577	255
Medium	131	264	605	324	131
Large	21	31	81	36	16
Extra Large	12	22	37	17	13

cohort, referred to as their "water group." Using the cohorts ensured that homes were only grouped with homes with roughly the same water needs. Importantly, a water group was assigned for all households in the experiment, including those receiving the control mailer. This allowed us to use the control group directly to analyze the effects of ranking, group performance, and other characteristics of the peer comparison. This is helpful because it allows us to consider what control households would have received as a peer comparison, had they been assigned to receive one.

Once households were assigned a water group from within their cohort, they were randomly assigned to one of the four experimental mailer groups in the study: a control mailer group (the "In-Sample Control" group) and three treatment mailer groups (the "Rank," "Team," and "Competitive Rank" treatment groups). The treatment and control groups are outlined in detail later in this section. Each of these groups received up to four Home Water Reports over the course of the experiment, but with different information in the treatment area, as outlined below.[24] This mailer was delivered to each household in the experiment every two months, by postal mail or email. Households in each experimental group got the same version of the mailer each time (in other words, a household assigned to the "Rank" treatment group received up to four "Rank" treatment mailers). A sample HWR is included in Appendix A.2, with the "treatment area" labeled.

A few things are worth noting about this setup. First, each of the households was linked

[24] Some homes did not receive all four mailers because of logistical issues or asynchronous timing of HWR delivery and water meter reads.

with unique water use outcomes—each water meter is associated with a single household unit. There is no complication from shared housing units with a single water bill. Second, all groups discussed so far received the Home Water Reports, which contained information about water use above and beyond what was randomly assigned to the treatment area. This information, which included a "WaterScore" driven by aggregate data on mean water use in the town, would likely have had an effect on water use independent of the experiment treatments. This is true for both the treatment mailer groups and the control mailer group. While this can be controlled for in the analysis to some extent, data was also collected from a neighboring town within the public utility's coverage area to serve as an additional control group (the "Out-of-Sample Control" group, discussed later). This data enables us to assess the efficacy of the mailers compared to the counterfactual of not receiving any mailers. Third, each individual household's water group was unique—just because household A was assigned a "water group" consisting of households B, C, D, and E, this did not (and in fact rarely) meant that household A appeared in B, C, D, or E's water group. Fourth, note that each household's water group consisted of homes in the same cohort but not necessarily in the same treatment group.

Treatments and Controls for In-Sample Households

Households in the sample were randomly assigned to receive one of four different mailer versions, which differed in the information displayed in the "Treatment Area" of the mailer, as labeled in Appendix A.2. Figure 1.1 displays the control and three treatment mailer versions.

Households assigned to the In-Sample Control group received the standard HWR, with a "Got water questions?" insert in the treatment area. Note that no information about the water group was transmitted to households in the In-Sample Control, nor was the household made aware that any comparison water group had been created.

Households in the Rank treatment group received an HWR with a ranked comparison in the treatment area. This treatment provided simple social rank information, with neutral

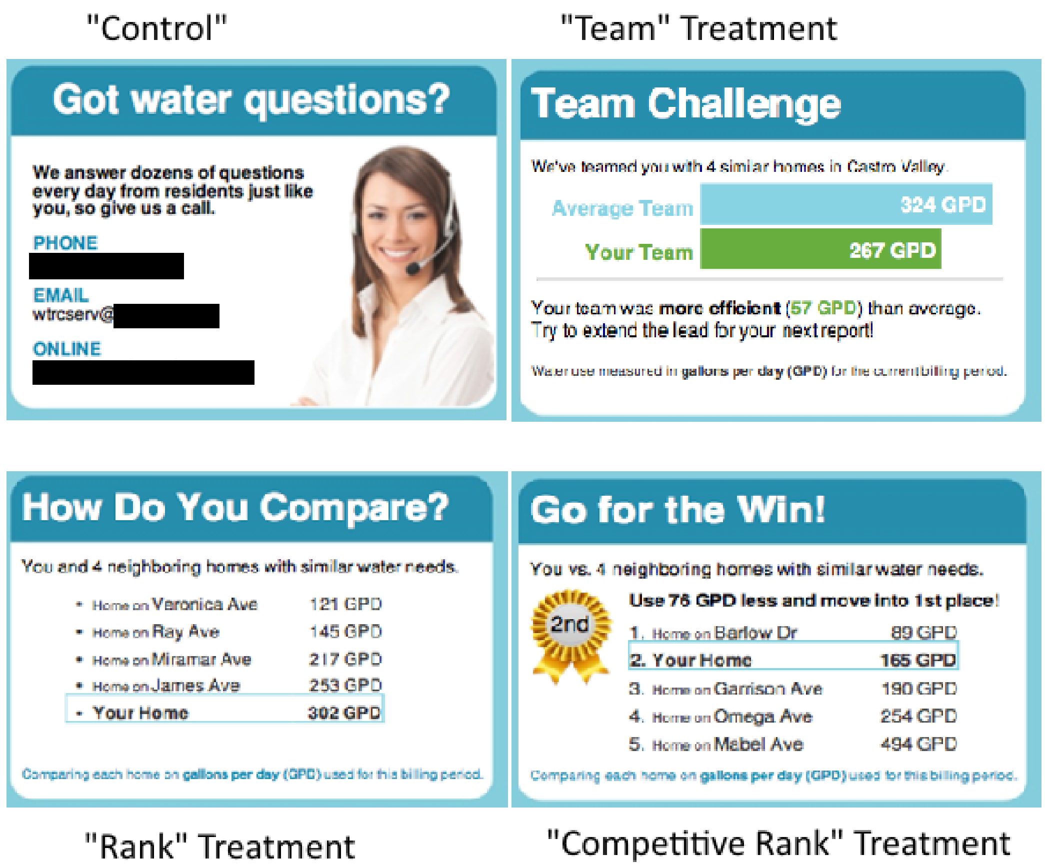


Figure 1.1: *Control and Treatment Mailer Versions*

framing.

Households in the Competitive Rank treatment group received an HWR with a competitively-framed ranked comparison in the treatment area. This treatment provided the same social rank information as the Rank treatment, but with a competitive frame (using "Go for the Win!" messaging and a ribbon icon), to assess the effect of priming a competitive instinct on behavior.

Households in the Team treatment group received an HWR with a cooperatively-framed "Team Challenge" in the treatment area. Note that the Team treatment did not include social rank or peer comparison information within the water group, but emphasized the group as a collective (and provided a comparison of the household's water group with other water groups). As a result, subjects in this treatment did not know the precise water use of the other homes in their water group, though they would have been able to deduce how their personal water use differed from that of their team by comparing the team average to their own usage (available elsewhere in the HWR).

Finally, since the In-Sample Control group shown above received mailers, it was important to have a second control group in the experiment that was not sent any mailers. For logistical and administrative reasons, it was not feasible to have this control group in the experimental location itself. Therefore, 2,880 households from the nearby Dingee pressure zone were used as an "Out-of-Sample Control" group for the experiment.[25]

Timeline

The experiment began in November 2012. The firm sent out the first mailers at the end of November, using October 2012 meter reads. The firm then sent three additional mailers, with the same treatment/control messaging, in January 2013 (based on December 2012 meter reads), March 2013 (based on February 2013 meter reads), and May 2013 (based on April 2013 meter reads). Households that had meter reads ouside of the four key meter read months (October 2012, December 2012, February 2013, or April 2013) did not receive a experimental mailer in the month that followed their read.

1.3.3 Data and Baseline Characteristics

We collected the data for this study from the firm, who obtained it through their partnership with the public utility. Two types of data were collected. First, we collected water use data for the households in the experiment, for the periods before and during the experiment. Second, the firm provided us with data on the characteristics of the households in the study, which they obtained both from the public utility and from independent data sources, including DataQuick.

[25]The Dingee pressure zone (code B5A) spreads across parts of the Berkeley/Piedmont/Oakland area. All homes in Dingee, with the exception of a small number of homes in the pressure zone that had received mailers as part of a prior firm pilot, were used as the Out-of-Sample control group.

Table 1.3: *Number of Households by Treatment*

Treatment	Number of Households	Number of Households receiving all four mailers
Control Group	1,308	1,091
Treatment #1: Rank	1,288	1,050
Treatment #2: Team	1,284	1,056
Treatment #3: Competitive Rank	1,300	1,068

Descriptive Data and Baseline Characteristics

We observe data from all 5,180 experimental households in Castro Valley, of which 4,265 received all four experimental mailers. Table 1.3 outlines the number of households in each treatment and control group, and the number of households in each group that received all four mailers. In addition, we observe data for 2,880 additional households from the neighboring Dingee area to serve as the Out-of-Sample Control. Appendix Table A.1 shows demographic information for both the In-Sample and Out-of-Sample households. The In-Sample households' demographics are presented broken up by treatment group. Note that the Out-of-Sample households differ from the In-Sample households in these observables, which is expected since they are in different areas. The Out-of-Sample homes and properties are, on average, slightly larger and older than the In-Sample homes, and used more water than the In-Sample homes in the 2012 months prior to the experiment.

Pre-Treatment Water Use Trends

Meter read technicians from the water utility measured water use every two months. Most meters used CCF units for water use (1 CCF = 100 cubic feet of water = 748 gallons), and the CCF reads were converted into a "gallons per day" (GPD) measure by the public utility. The mean water use in the In-Sample area (as well as in the Out-of-Sample Control area) prior to the experiment, measured in GPD, is visible in Appendix Table A.1. Additionally, Panel (A.3.1) in Appendix A.3 provides a graphic of the water use trends in the two areas prior to the experiment. Notice the key role that seasonality plays in water use; water use is higher in the summer than in the winter. Because of this, we used month fixed effects in certain specifications to control for seasonal trends.

Note that while mailers were sent on the same date for all households in each mailing cycle, households did not have meter reads on the same date. As a result, there is some variance in how many days a given household was treated by a single mailer. For example, some households would have received the mailer in the week before their next meter read while others received it in the week after their previous meter read. This is not an uncommon issue, having appeared in similar experiments using read-based mailers, including Allcott (2011). Successful randomization deals with this issue to some extent, in that there is no correlation between treatment and meter read cycle. This means that, on average, each condition's households were "treated" during a given mailer cycle for the same fraction of the post-mailing period.

Randomization Check

Some recent research questions the need for randomization checks in experiments (Mutz and Pemantle (2011), for example). However, in this instance randomization checks are warranted for two main reasons. First, the randomization process itself was conducted by the firm and not the researcher. Though the firm has a track record of experimentation and a strong background in randomization procedures, a check is needed to ensure that there were no systematic errors in randomization. Second, some households were dropped after randomization but prior to study implementation. In particular, 355 households did not receive a mailer despite being assigned to one of the treatment or control groups, for logistical reasons (the subject moved from the property, the address was not verified, etc.).

To test the balance of the samples on these observed demographic characteristics, we run a regression of the various demographic characteristics (written as y_i below) on dummy variables for the three treatment groups (written as T_m with m ranging from 1-3 for the three treatment groups, below), omitting the In-Sample Control. We also compute f-test statistics to determine joint significance. The econometric model is as follows:

$$y_i = \beta_0 + \sum_{m=1}^{3} \beta_m (T_m)_i + \varepsilon \tag{1.1}$$

Appendix Table A.2 presents the results from these regressions. None of the f-statistics and associated p-values suggest joint significance for the coefficients. We are therefore confident that the randomization resulted in balanced treatment and control groups.

Handling Outliers

The primary outcome measure in a given water read cycle, GPD, had occasional extreme values. We remove outliers from the analysis on both the high and low ends. First, households occasionally register a GPD of zero for a given read period. This is generally because household members are either not at home during the read period, or because their water use is so low that it fails to register on the water meter. In all household water use data (8,060 households, including the 2,880 Out-of-Sample Control households), only 135 households had a zero GPD reading for at least one meter read period after October 2011 (the relevant period for the difference-in-differences analysis used in this paper), and 90 households in the In-Sample area had at least one zero reading during the experimental period (December 2012 - June 2013). To prevent these low values from influencing the results, we conduct our analyses without these zero GPD observations, depending on the periods being analyzed in a given specification.[26]

Second, there were a few meter reads that were far above normal values. One observation in particular was dramatically in excess of normal levels (over 120,000 GPD for a single household read—the median value of average household water use in 2011 was 193.68 GPD). The utility identified these high reads as meter malfunctions or abnormalities. To deal with such outliers on the upper tail of the distribution, we use 5,000 GPD as a cutoff for a single

[26]For any analysis that used only post-experiment, In-Sample group data (means comparisons across In-Sample groups and the rank effects analysis, for example), only those GPD reads of zero after the experiment was initiated were excluded. However, for difference-in-differences analysis using both In-Sample and Out-of-Sample data from the year prior to treatment as a "pre" period, any household with a zero GPD reading in any single read after October 2011 was excluded.

meter read to define excluded outliers. In total, three households in the data had a read in excess of 5,000 GPD after October 2011, and two households had a read above this threshold during the experimental period.

1.4 Empirical Methods

In this experiment, subjects received multiple treatment mailers. The information in the treatment area of the mailers differed by treatment group, as did the information displayed in the non-treatment sections of the mailers. As a result, we must use a variety of econometric methods to analyze the experiment and its effects. Also, the distinction between the In-Sample and Out-of-Sample Control groups is important, as comparisons between the treatments and the two control groups require different interpretations. In this section, we outline the econometric strategies used for answering different research questions in the data.

1.4.1 Average Treatment Effects: Difference-in-Differences, Matching, and Regression

We estimate the average treatment effect in three ways. First, we use the Out-of-Sample Control group to determine overall mailer effects in the short and long run, using a difference-in-differences approach. Second, we test the robustness of the difference-in-differences results using nearest-neighbor matching. Finally, to assess the impact of the different treatment mailer versions on household water use behavior, we use regressions that compare means across treatments while controlling for relevant variables, then disaggregate the analysis using past water use.

Difference-in-Differences (In-Sample vs. Out-of-Sample)

The first question we seek to address is whether the HWR mailers, including the control mailer, influenced water use. To do this, we use a neighboring pressure zone as an Out-of-

Sample control and perform two types of analysis. First, we use a difference-in-differences approach.[27] The "parallel trends" assumption needed for difference-in-differences analysis seems credible given the data from the pre-treatment period. The pre-experiment mean water use in the two areas is visible in the panels in Appendix A.3, which provide visuals to justify the parallel trends assumptions.

As Panel (A.3.1) in Appendix A.3 shows, one concern may be that the gap in mean water use between the In-Sample and Out-of-Sample areas was larger in the summer than in the winter. Therefore, when using the difference-in-differences approach, we looked at the matching month or months for each household in the year before and the year after the treatment to control for the seasonal water use differences between the areas. We first restrict the analysis to the first meter read after the initial mailer (December 2012), comparing it to the read from the same month a year prior to treatment (December 2011). Using this approach enables us to handle differential seasonality and attain something closer to the "parallel trends" needed for difference-in-differences analysis. Panel (A.3.2) in Appendix A.3 depicts this visually. The econometric specification is as follows:

$$GPD_i = \beta_0 + [\sum_{m=1}^{4} \beta_m (T_m)_i * (Post)] + [\sum_{m=5}^{8} \beta_m (T_m)_i] + \beta_9 (Post) + \varepsilon \tag{1.2}$$

In this specification, "Post" is a dummy variable for whether or not the reading was from December 2012. The β_1,β_2,β_3, and β_4 coefficients serve as difference-in-difference estimators of the causal impact of the mailers on water use. Note that there are four treatment dummies here (In-Sample Control, Rank, Competitive Rank, and Team), with the Out-of-Sample Control the omitted group. Furthermore, no household characteristic controls are used, as these variables did not vary within a household over time in the data.

We then repeat this approach using the mean water use in each household in all relevant pre- and post-experiment periods.[28] Specifically, we use the mean household water

[27]This is necessary because the Out-of-Sample and In-Sample areas differ in observable characteristics and water use patterns.

[28]Recall that most households received all four treatment mailers.

consumption in the December 2012, February 2013, April 2013, and June 2013 meter reads as the post-experiment water use outcome. We then gather data on the water use for these households in the matching months pre-experiment (so, from the December 2011, February 2012, April 2012, and June 2012 reads), and compute a mean for each household in the four pre-experiment reads. By doing this, we obtain a pre- and post-experiment mean that aggregates multiple meter reads.[29] We then run the above regression specification for a difference-in-differences estimator using the overall mean water use, to approximate an overall effect of the mailers. This approach helps us deal with the seasonality issue across the treated and control areas, and strengthens the justification for the parallel trends assumption (as visible for this specification in Panel (A.3.3) in Appendix A.3). Furthermore, it acknowledges the inherent issues with serial correlation in difference-in-differences estimators raised by Bertrand *et al.* (2004), improving our faith in the standard errors of our estimates.[30]

Matching Estimators (In-Sample vs. Out-of-Sample)

To check the robustness of the above difference-in-differences results, we compute average treatment effects using a matching framework. Specifically, we use nearest-neighbor matching to match individual households in the In-Sample area (who all received mailers) to households in the Out-of-Sample area (who did not receive mailers). The matching is based on the water use of the households in the four water reads prior to the experiment, and three household characteristics that are plausibly linked to water use (home size, lot size, and exact matching to the number of occupants in the home). The outcome variables in this analysis are water use in the first period after mailer initiation (to estimate the short-run effect) and mean water use over the four periods following mailer initiation (to estimate the

[29] When all four reads were not available either pre- or post-experiment for a given household, the mean for any matching pre/post months that were available was used instead.

[30] The authors write, in reference to the serial correlation problem in many difference-in-differences estimates, "collapsing the data into pre- and post-periods produce consistent standard errors, even when the number of states is small." (see Bertrand *et al.* (2004, p. 274))

long-run effect).[31] When appropriate, these matching estimators use the bias adjustment procedure outlined in Abadie and Imbens (2011) and Abadie *et al.* (2004), which adjusts for differences in matches in finite samples. Also, both Euclidian distance and Mahalanobis distance are used as matching metrics, though in the paper we present the matching results using Mahalanobis distance only.[32]

Regression for Means Comparison (In-Sample)

To determine whether mailer version influenced water conservation behavior, we use regressions that compare mean water use across mailers after the initiation of the treatment, in two ways. First, we compare mean water use in the first period (from the meter read following receipt of the first mailer) across treatment groups. Second, to provide an estimate of the long-run differences in water use across treatment mailers, we compare mean water use for all post-treatment periods across conditions.

The econometric specification for this comparison is a regression of water use by the household (in gallons per day) in the relevant periods on dummy variables for the three treatments, and is visible below:

$$GPD_i = \beta_0 + \sum_{m=1}^{3} \beta_m (T_m)_i + \varepsilon \tag{1.3}$$

We also run these regressions with controls for home characteristics (lot size, home square footage, and the number of bathrooms, bundled below as ϱ_i). This specification is shown below:

$$GPD_i = \beta_0 + \sum_{m=1}^{3} \beta_m (T_m)_i + \varrho_i + \varepsilon \tag{1.4}$$

Importantly, we disaggregate the analyses to determine whether the average treatment

[31] For this analysis, we exclude homes that were missing at least one meter read during the four key pre and post periods. This reduced the number of households in the analysis from 4,908 to 4,109 for the various in-sample households, and from 1,668 to 1,426 for the out-of-sample households.

[32] The match results based on Euclidian distance show similar effects, and will be made available through the online appendix.

effect differs across conditions based on past water use. Specifically, we classify households as being "low" or "high" water users using data on water use in the pre-experiment reads in 2012. Low-use households are defined as those in the bottom third of water use within each irrigable area category, and high-use households are defined as those in the top third within each irrigable area category. By assessing water use within the irrigable area classifications, we are able to control for the differences in water needs based on property size, which allows for large homes with efficient residents to still be classified as "low" users.

This diversity of approaches helps us address theories about the differential effect of framing on household response to peer information, capturing heterogeneities in treatment effects that a simple average treatment effect will miss. The critical specifications center on identifying if certain mailer versions were more or less effective for households with "high" or "low" water use, pre-experiment. This allows us to test the hypotheses around potential "boomerang" or "what-the-hell" effects in social comparison messaging.

1.4.2 Ranking Effects

In the Rank and Competitive Rank treatments, which both displayed social rank information relative to four peers, each possible rank position can be viewed as a distinct treatment. In other words, a "first place" Competitive Rank mailer may result in a different behavioral response than a "last place" Competitive Rank mailer. This feature of the experimental setup allows us to explore and test theories about social rank and its influence on behavioral response. We do this using two different econometric strategies.

Restrict Focus to Last/First Place Mailers Only

First, we treat each mailer and the household's water use in the period that follows as a distinct treatment/outcome pair. This requires a model of behavioral response whereby a household's behavior in the period following a mailer is a direct response to the content of that mailer and is independent of the content of previous mailers. From the perspective of maintaining randomization, this is not an issue with the first mailer and subsequent behavior. However, since we use multiple mailers per household in the analysis, this does threaten our identification by moving away from pure randomization. This is because the content of previous mailers may have influenced household response to subsequent mailers.

There are precedents for this approach to assessing the impact of multiple treatments in existing research. For example, Doherty and Adler (2014) argue that mailer effects in a political campaign context are short-lived. The authors suggest that individual-level responses can be considered in the period immediately following a given mailer, as timing may be more important to outcomes than mailer quantity. Additionally, Allcott and Rogers (2014) find evidence of cycles of significant backsliding in the weeks immediately following social information mailer receipt, using data from Opower's Home Energy Reports. A similar sort of backsliding here could lead to near-total decay of mailer effects by the end of a single period post-mailer.

One could also justify this approach, as Bertrand *et al.* (2010) do, using the behavioral concepts of "System I" and "System II" thinking (Stanovich and West, 2000). That is, the receipt of a given mailer (and the social information contained in the mailer) can have two effects. First, in the short run it can cause an intuitive, System I response in the household, whereby the specific information in the mailer has an immediate effect on behavior. Second, in the long run they may invoke a System II impact, whereby the mailer causes deliberative changes in behavior (such as replacing household fixtures). It is plausible that in our data, the short-run, System I response would be more visible than the long-run, System II impacts of any previous mailers.

While we feel this analytical technique is justifiable in this case, we attempt to control for potential biases from repeat mailer exposure by using fixed effects for the number of mailers seen prior to the read in question. These controls do not significantly change our results. We feel this justifies our identification assumption that the impact of a given mailer on behavior is primarily restricted to the period immediately following that mailer's receipt, and is independent of prior messaging.

To assess the rank effects econometrically, we look at all mailer/outcome pairs for households that finished in “last place” in that specific mailer, and regress household water use on the treatments. We only do this for the In-Sample Control, Rank, and Competitive Rank treatments, as the Team treatment does not have any visible ranking. Note that households in the In-Sample Control group are assigned to water groups, but information about their group is never displayed to them. Consequently, “last place” homes in the two treatment groups can be compared to would-be “last place” homes in the In-Sample Control group. The In-Sample Control group is the omitted group in the regression, leaving regression coefficients that represent treatment effects of Rank and Competitive Rank mailers for “last place” homes. We then repeat this analysis for “first place” households. Since there are multiple observations for each household in the sample in this analysis, we cluster standard errors at the household level. The specification is shown below.

$$GPD_{ijk} = \beta_0 + \beta_1(T_{Rank})_i + \beta_2(T_{CompRank})_i + \beta_3(MailerGPD_{ijk}) + \varrho_i + \delta_j + \gamma_{ijk} + \rho_k + \varepsilon_{ijk} \tag{1.5}$$

Note that the specification includes controls for household water use displayed (in gallons per day) in the mailer ($MailerGPD_{ijk}$), household demographics (lot size, home size, and bathrooms, captured by ϱ_i), month fixed effects (δ_j), WaterScore fixed effects (γ_{ijk}), and fixed effects for the number of mailers seen prior to the observation mailer (ρ_k).

Rank Effects Amongst Middle Third of Water Users

A second approach to evaluating rank effects is to restrict attention to homes in the middle third of water users pre-experiment, whose water use was around the mean given their irrigable area. Due to the random assignment of water groups, some of these households had a water group composed of relatively low (or high) water users. Therefore, there is variation in rank position amongst these homes that is not a function of their actual water use behavior. We exploit this variation and run the following specification, which uses data from all mailers received by households in the middle third of water users pre-experiment, and includes interaction effects between rank position (1st, 2nd, 3rd, 4th, or 5th) and treatment

(Control, Rank, Competitive Rank):

$$GPD_{ijk} = \beta_0 + [\sum_{m=1}^{3}(\sum_{n=1}^{5}\beta_{m,n}(Position_n)_{ijk} * (T_m)_i)] + \beta_{15}(MailerGPD_{ijk}) + \varrho_i + \delta_j + \gamma_{ijk} + \rho_k + \varepsilon \tag{1.6}$$

The 14 interaction terms will reveal whether or not there is a differential impact of rank position based on mailer version, which will help us identify any evidence for a "last place effect" or "first place effect" in the Rank and Competitive Rank treatments (note that all In-Sample Control households are unaware of their social rank position and therefore serve as an effective control group). Again, this specification controls for household water use displayed in the mailer ($MailerGPD_{ijk}$), lot size, home size, and bathrooms (ϱ_i), month fixed effects (δ_j), WaterScore fixed effects (γ_{ijk}), and fixed effects for the number of mailers seen prior to the observation mailer (ρ_k). Standard errors are again clustered at the household level.

1.5 Results

1.5.1 Overall Mailer Effects: Difference-in-Differences and Matching

We begin by using the Out-of-Sample Control group to assess the impact of the four experimental mailers (the In-Sample Control, Rank, Team, or Competitive Rank treatments) on water use. The short-run results, which focus on the read following the first mailer (December 2012) compared to the same month in the previous year, suggest that the experimental mailers did not have a significant effect on water use in the short run. These results are visible in Appendix Table A.3, in both linear and log forms, and in the Tables 1.4 and 1.5 here. We also compute matching estimators as described in section 1.4.1. The matching results suggest a greater short-run decrease in water use than the difference-in-differences estimators (particularly when matching was done using only one match). However, when we use four matches instead of one, the effect sizes are smaller. Taken together, we conclude from these results that the mailers had minimal impact in the short run.

Table 1.4: *Short-Run ATE Estimates (for the first period post-mailer initiation, December 2012)*

Mailer Version	Difference-in-Differences	Matching (1 Match)	Matching (4 Matches, bias-adjusted)
In-Sample Control	1.65 GPD (0.81%)	-21.52 GPD *** (-6.09%) ***	-8.49 GPD * (1.41%)
"Rank"	4.24 GPD (1.36%)	-19.84 GPD *** (-4.80%) **	-7.41 GPD * (1.55%)
"Team"	4.26 GPD (0.24%)	-21.22 GPD *** (-6.38%) ***	-6.92 GPD (1.60%)
"Competitive Rank"	2.11 GPD (0.41%)	-22.62 GPD *** (-6.46%) ***	-15.45 GPD *** (-0.55%)

For convenience, we present the short-run ATE estimates from the difference-in-differences and matching analyses, in both level and log form, in Table 1.4.

The results for the entire experimental period, however, suggest that the different mailers reduced water use by 13-17 gallons per day, or around 3%. The reduction in water use in the difference-in-differences regression was statistically significant at the 95% level for the In-Sample Control (16.17 GPD) and Rank (16.28 GPD) mailers, and at the 90% level for the Team mailer (14.62 GPD) and the Competitive Rank mailer (13.44 GPD). However, none of the log specifications are significant (though the estimates are all in the 3-5% range). Appendix Table A.4 shows these results. To verify the difference-in-differences estimates, we compute average treatment effects using matching as described in section 1.4.1. The matching results are similar in magnitude to the difference-in-differences results. We present the overall ATE estimates from the difference-in-differences and matching analyses, in both level and log forms, in Table 1.5.

These results suggest that the initial mailer with social information failed to spur significant behavior change in the short run, but over time the mailers did influence behavior. It is possible that the effects observed here are artifacts of a breakdown in the parallel trends assumption underlying the analysis. However, given the past data on the trends in water use in the In-Sample and Out-of-Sample areas, we are relatively confident that the observed patterns capture genuine conservation efforts on the aggregate in response

Table 1.5: *Overall ATE Estimates (Estimates for all post-mailer initiation reads, December 2012-June 2013)*

Mailer Version	Difference-in-Differences	Matching (1 Match)	Matching (4 Matches, bias-adjusted)
In-Sample Control	-16.17 GPD ** (-3.36%)	-27.87 GPD *** (-7.84%) ***	-12.09 GPD *** (0.58%)
"Rank"	-16.28 GPD ** (-3.35%)	-31.48 GPD *** (-8.08%) ***	-17.87 GPD *** (-0.66%)
"Team"	-14.62 GPD * (-4.59%)	-23.61 GPD *** (-6.00%) ***	-9.83 GPD *** (1.61%)
"Competitive Rank"	-13.44 GPD * (-3.23%)	-24.14 GPD *** (-6.70%) ***	-9.56 GPD *** (1.03%)

to the mailers.

The most interesting result here is that the Competitive Rank mailer was the least effective across specifications. This hints at possible underlying differences in how individuals responded to the mailers, particularly the Competitive Rank mailer. We explore explanations for these differential impacts in the next section.

1.5.2 Across Mailer Differences: Means Comparisons with Regression

Having found evidence that the mailers do influence behavior, we now set aside the Out-of-Sample Control, and look only at the households that did receive mailers, to explore differences in behavioral response across mailer versions.

Aggregate Means Comparison Regressions

Appendix Tables A.5 and A.6, along with Figures 1.2 and 1.3, present the results of the regressions of water use on treatments for all in-sample groups. The goal of this analysis is to estimate the average treatment effect of each version of the mailer relative to the control mailer, which featured no social information except for the "WaterScore" on the left hand side of the mailer. Appendix Table A.5 and Figure 1.2 present the results from regressions using water use in the first period following experiment initiation as the outcome variable. Appendix Table A.6 and Figure 1.3 present the results from regressions using mean water

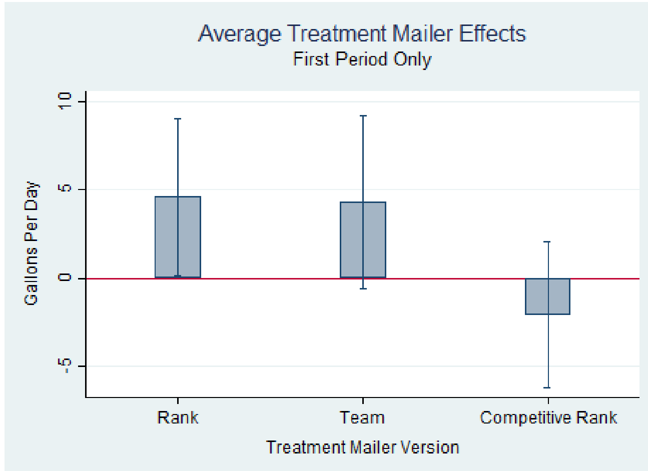


Figure 1.2: *Short-Run ATE (Dec. 2012 Only) - SE Marked*

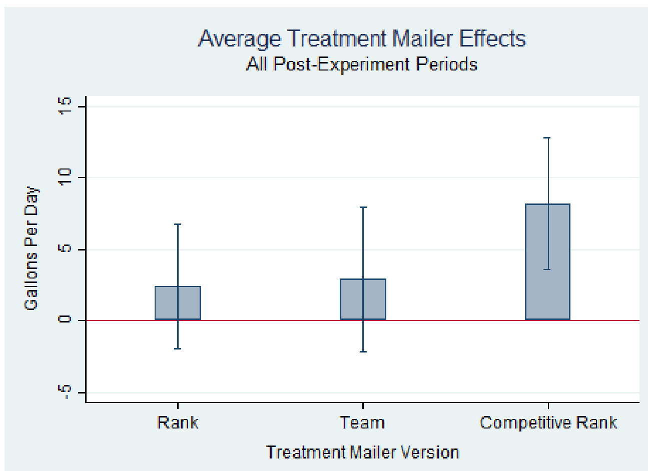


Figure 1.3: *Overall ATE (All Post-Experiment Periods) - SE Marked*

use in all periods following experiment initiation as the outcome variable.

The results show no strong evidence of a differential impact across mailers in the short run or overall. The only statistically significant result (at the 90% level) is that the Competitive Rank mailer performed worst, increasing water use by 8.19 GPD relative to the control mailer overall (Appendix Table A.6, specification (2)). The similarity in effect across mailer versions is not surprising, since this analysis treats receipt of any version of a given mailer as part of the same treatment, whether or not you performed well or poorly in the displayed peer comparison. In other words, a household receiving a Competitive Rank mailer and finding themselves in "first place" in the ranking is, in this analysis, grouped with a household receiving a Competitive Rank mailer and finding themselves in "last place." It is likely that these two types of households will have very different responses to the Competitive Rank mailer. More analysis is therefore needed to understand possible heterogeneous effects of the mailers on household behavior (and follows in section 1.5.3).

However, this is still an important result from a policy perspective. No mailer version is inherently better than the other, in terms of its average effects. This suggests that a policymaker seeking to make a blanket decision on which form of messaging to use in a social information mailer across a large population cannot expect one type of messaging to work better. Instead, a targeted approach that considers disaggregated treatment effects may prove most effective in influencing behavior.

Disaggregation by Past Water Use

We next repeat the analysis, but disaggregate based on a key, visible covariate—past water use. By doing so we can determine if certain mailer versions were more effective for households with high or low water use pre-experiment.[33] Appendix Tables A.7-A.14 show the output from these regressions, while Figure 1.4, Figure 1.5, and Appendix A.4 provide visuals, as noted below.

The results suggest that the Rank treatment increases water use amongst households with low water use prior to the experiment, relative to the control mailer. The effect is mostly from the impact of the mailers in the first post-mailer period, where the Rank treatment increased water use by 12.02 GPD relative to the control mailer, statistically significant at the 95% level (Appendix Table A.7, specification (3), and Appendix A.4, Panel (A.4.1)). It is instructive to compare the Rank and Competitive Rank mailers directly as well, since the only difference between these mailer versions was the framing around social rank information. When compared directly with the Rank mailer, the Competitive Rank mailer is

[33]High and low water users are defined as being in the top/bottom third of water users within their irrigable area category in the pre-experiment period in 2012. See section 1.4.1 for more.

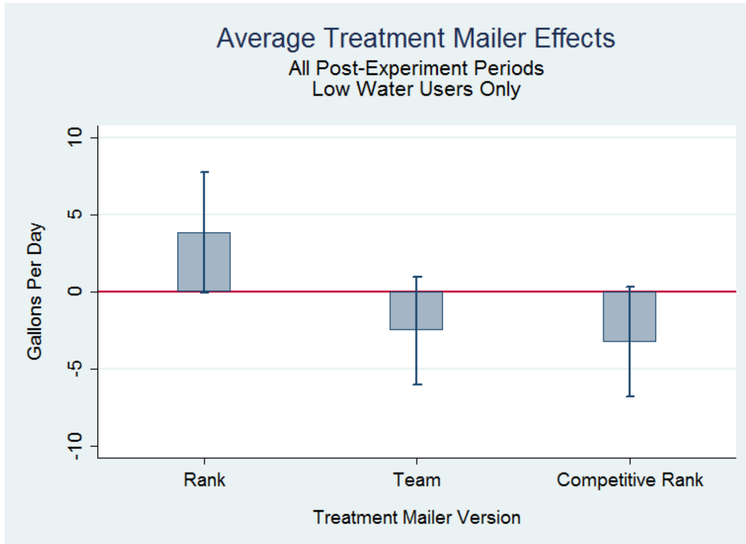


Figure 1.4: *Overall ATE (Low Water Users Only) - SE Marked*

associated with 14.22 GPD lower household water use in the first period—10.3% less than the Rank mailer (Appendix Table A.8, specifications (2) and (4), and Appendix A.4, Panel (A.4.1)). When looking at the mean water use during all periods following the initiation of the experiment, however, the detrimental effect of the Rank mailer relative to the control mailer is smaller and not statistically significant (3.85 GPD more than the control, around 3.0% higher water use when a log-level regression is used, visible in Appendix Table A.9, specifications (2) and (4), and Figure 1.4). However, the difference between the Rank and Competitive Rank mailers remains significant at the 90% level, with the Competitive Rank mailer associated with 7.15 GPD lower household water use than the Rank mailer over the entire experimental period (visible in Appendix Table A.10, specification (2)).

This is a notable result—there is evidence of a "boomerang effect" for low water use households from rank information, but one that is counteracted by a competitive frame. Note that this effect comes solely from the peer comparison, and not the other social information on the mailer, which is controlled for in the regression. One possible explanation for this finding is that the Rank treatment's neutral messaging does not provide sufficient incentive for efficient households to continue conservation efforts. The Competitive Rank treatment mailer provided peer comparison and social rank as well, but did so with an

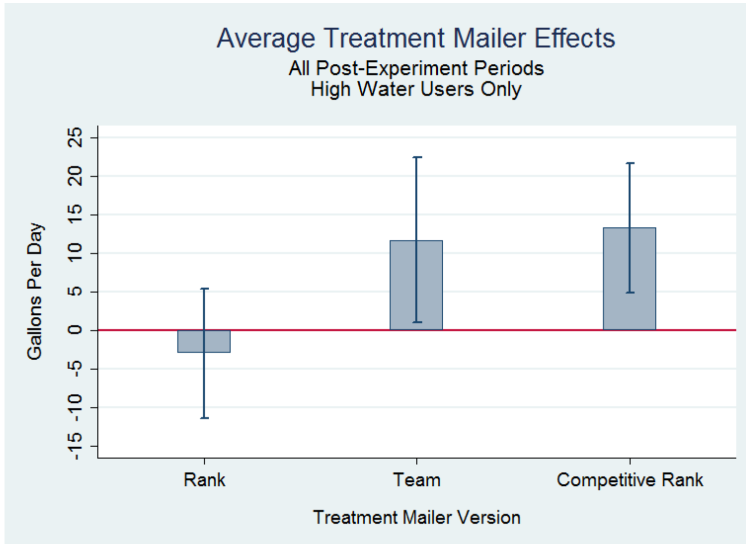


Figure 1.5: *Overall ATE (High Water Users Only) - SE Marked*

added competitive motivation, which arguably prevents the boomerang effect observed for households receiving the Rank mailer.

Meanwhile for households with high levels of water use pre-treatment, the treatment effects are different, and less statistically compelling. As Appendix Tables A.11 and A.12 demonstrate, the mailers were similarly effective in the short run. While the Team treatment performed slightly worse than the other two treatments, this difference was not statistically significant (Appendix Table A.11, specifications (3) and (6), and Appendix A.4, Panel (A.4.2)). When we look at the mean water use in all periods following the first mailer, in Appendix Tables A.13 and A.14, we see that the Competitive Rank mailer performs worse than the other mailers, increasing mean water use by 13.28 GPD relative to control mailer (Appendix Table A.13, specification (2), and Figure 1.5) and by 15.89 GPD relative to the Rank mailer (Appendix Table A.14, specification (2)). The second result is statistically significant at the 90% level, and equates to roughly 4.0% higher water use than the Rank mailer when using the log-level specification (Appendix Table A.14, specification (4)).

This is suggestive of a competitive framing effect for high water use households that is the exact opposite of that for low water use households—while competitive framing of rank information had a positive effect on low water use households (preventing a boomerang effect), it seemed to increase water use in high water use households. This could be because it is demotivating to perform poorly in a competitive comparison with your peers. Note that the higher water use relative to the control group was not observed in the Rank treatment (Appendix Table A.13, specification (2))—the competitive framing seems to be the key element.

1.5.3 Ranking Effects

In assessing the effect of specific rankings, we focus on first and last place in particular. We begin by restricting analysis to the following mailers and subsequent outcomes: 1) households receiving first/last place rank messaging in the Rank and Competitive Rank treatments; and 2) households receiving the Control mailer who "would have" ranked in first/last had they been shown a displayed rank. We use regressions to estimate the effect of displayed "first" and "last" place messaging on behavior following mailer receipt using data from all experimental mailers, and we cluster standard errors at the household level. The full results on ranking effects are in Appendix Tables A.15 and A.16.

The main result in this analysis is the existence of a persistent and strong "last place effect" for households in the Competitive Rank treatment (Appendix Table A.15). Specifically, households ranked last in the Competitive Rank treatment show higher post-mailer water use than households in the In-Sample Control who would have been in last place had they seen their position (17.85 GPD more water use than the In-Sample Control mailer, visible in Appendix Table A.15, specification (4)). This effect is also significant in comparison to the last-placed individuals in the Rank treatment.[34]

The results suggest that priming a sense of competition makes social rank information demotivating for people who perform worst in the displayed rank. This is especially interesting because the Competitive Rank treatment does not seek to prime negative thoughts about poor performance in the household—it actually encourages a last place household to

[34] The Rank treatment seems to cause a smaller "last place effect" of its own relative to the control when a log-level specification is used, however.

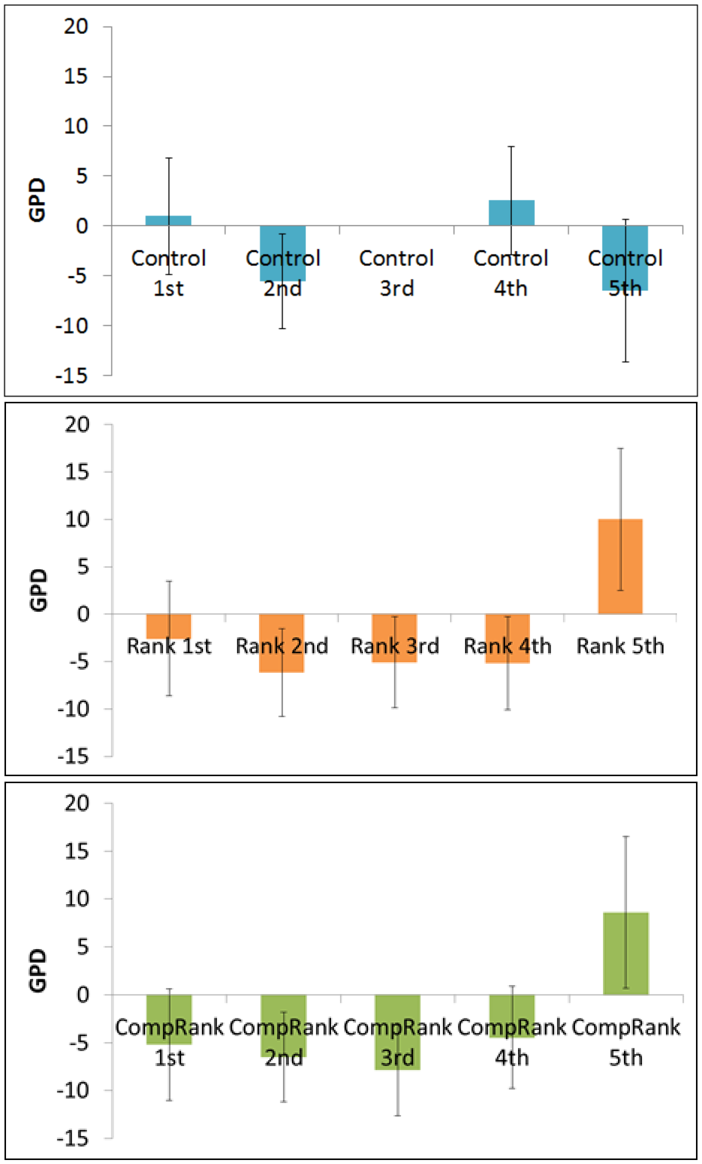


Figure 1.6: *Coefficients from Interactions of Treatment and Rank Position: Middle Third of Water Users*

Note: This figure shows the results of the regression based on equation (1.6), reported in Appendix Table A.17, specification (3). Clustered standard errors are marked with bars.

improve in an effort to attain 4th place.

The evidence for a comparable "first place effect" is not as compelling. As specification (4) in Appendix Table A.16 shows, the visible "first place effect" in a simple specification without controls disappears with the inclusion of controls.

For robustness, we use the second approach outlined in section 1.4.2, restricting analysis to only those homes in the middle third of water users. We estimate the effect of rank here by interacting treatment and rank to determine if there was a differential response to rank position by treatment. Appendix Table A.17 presents the results of this regression, and Figure 1.6 provides a visual depiction of the coefficients on the interaction terms by treatment and position (because it was necessary to omit a coefficient, Control households in 3rd position are omitted to generate the figure). While not statistically significant, the clear message from this analysis is that the Rank and Competitive Rank treatments seem to consistently drive up water use for households in last place.

When these results are coupled with the earlier results showing that the Competitive Rank mailer performed worst on aggregate, a clear story emerges. The competitive framing discourages high water users, particularly those individuals who find themselves in "last place" in a displayed rank. These individuals perform worse than they would have had they not seen the rank information. Simultaneously, the competitive frame has a smaller positive impact on low water users. However, the detrimental effects of the competitive frame (on the high water users) outweigh any positive impacts (on the low water users), meaning that on aggregate the Competitive Rank mailer performs worst of all mailer versions used.

1.6 Discussion and Conclusions

Our experiment provides insights on some important underlying drivers of behavioral response to social rank and peer comparison. First, the experiment replicates existing work on social information by showing that mailers using this information can reduce water use (by roughly 13-17 gallons per day in this study). Overall, the experiment finds that the different frames used in the mailers (neutral, competitive, and cooperative) had slightly different effects on water use, with the competitively-framed rank mailer performing marginally worse. However, this aggregate comparison of mailers masks more interesting results on the underlying mechanisms of rank and response.

The most robust results come from the disaggregated analysis of treatment effects, and the analysis of specific rank effects. The analysis shows that the display of a neutrally-framed peer comparison with four similar homes caused a "boomerang effect" in water-efficient households, increasing the households' water use relative to the control. Interestingly, this boomerang effect was eliminated by the inclusion of a competitive frame. Though the results do not support the existence of a motivating "first place effect," the data overall suggests that high achievers thrive (or, at least, do not struggle) in competitive settings, and may need competition to avoid boomerang effects from explicit rank information.

However, the competitive framing of rank information had large demotivating effects on water-inefficient households. These households responded poorly to the competitively-framed rank information, more than offsetting the beneficial effects of the competitive framing on high achievers. Furthermore, it appears that rank effects play a significant role here as well. The results show that households who finish in "last place" in a competitively-framed rank comparison were demotivated, increasing their water use relative to households receiving the control mailer or the neutrally-framed rank mailer. Interestingly, the latter comparison implies that the demotivation effect observed is primarily driven by the competitive frame (rather than by the low ranking).

These results have direct implications for the competing theories related to rank and response. This experiment finds that in competitive settings, the theories on motivation and self-efficacy seem more consistent with observed behavior, with top performers holding steady while poor performers worsen. This supports the existence of a "what-the-hell effect" (Polivy and Herman, 1985), and also is consistent with the oppositional reactions to peer information found by Beshears *et al.* (2015) in the savings context. Furthermore, the finding that rank information offsets the boomerang effect for top performers is consistent with Garcia *et al.* (2006). However, without the competitive frame, neutral rank information seems to encourage a behavioral response that is more in line with social norms and social comparison theories, with "boomerang effects" for top performers the clearest result. This provides some structure to existing theories on rank and response, and suggests that the framing of peer comparisons and social rank is key to their success or failure.

The implication of these findings for public policymakers and "nudgers" seeking to promote conservation behavior is mixed. On the one hand, there seem to be benefits from social information overall. The mailers did influence behavior on the aggregate. However, the experiment also reveals some potential pitfalls of social information, namely that it can demotivate poor performers in a way that has detrimental effects on their efforts to conserve. This leads to important follow up questions. What types of social information are best to motivate those who are performing poorly? Why does a competitive frame prevent backsliding for top performers, and to what extent is this context-dependent? Further research is needed to better understand the observed effects and what forms of social messaging are needed to negate those effects.

Another important takeaway from the results in this paper is that behavioral phenomena in real-world settings can be very complex. Based on the results, it seems plausible that multiple behavioral forces are at play in this context—social pressure, boomerang effects, demotivation, and last place effects, to name a few. Furthermore, these forces seem to influence different subgroups of the population in different ways. These results suggest that future research in behavioral science more broadly should provide more than the simple average treatment effect, and explore other measures of impact (most notably heterogeneous treatment effects). Such estimates are not only easy to compute—they also significantly improve our understanding of how people behave and maximize the knowledge benefits from randomized evaluations.

On the whole, experiments of this form offer a compelling way to develop public policy around pro-social behavior change.[35] In the context of water use, exploring messaging and developing innovative ways to convey information to households can increase the salience of water use, cost, and environmental impact, and lead to changes in aggregate water use outcomes. Follow up research could extend this work in a number of ways. First, the "last place effect" observed here should be tested in a randomized setting with a larger sample size. Second, researchers should investigate whether increasing the salience of water cost influences behavior, and how to frame messaging around water bills to

[35]For a summary of field work in this broader area, see Kraft-Todd *et al.* (2015).

increase its significance to households. Third, future research needs to explore how low-cost messaging can be used to promote major household behavior change. In some ways, water conservation suffers from an "energy paradox" as outlined in Jaffe and Stavins (1994).[36] That is, home repairs that reduce water use may have significant long-run financial benefits for households, but the upfront cost and mental effort may prevent households from pursuing them. While behavioral nudges using social information might provide cost-effective and environmentally-significant savings, further work should explore the use of behavioral science to influence decision making on major, water-reducing investments in the home, like the replacement of water-hungry appliances. Perhaps influencing these "big" behaviors will be the key to locking in sustainable reductions in household water use.

[36]Note: Allcott and Taubinsky extend this "energy paradox" to electricity conservation as the "lightbulb paradox" in Allcott and Taubinsky (2014).

Chapter 2

Cheap Promises: Evidence from Loan Repayment Pledges in an Online Experiment

2.1 Introduction

A growing body of research in behavioral economics suggests that people find it challenging to stick to financial commitments, be it saving more, spending less, or repaying loans. The underlying reasons identified in the literature for these tendencies include limited attention, status quo bias, and time inconsistency (Banerjee and Mullainathan, 2008; Soman *et al.*, 2005; Samuelson and Zeckhauser, 1988; Karlan *et al.*, 2010). In light of this, researchers have studied how altering choice environments and decision-making contexts can help people overcome their biases and make welfare-improving decisions. For example, in an experiment in the Philippines, Ashraf *et al.* (2006) showed that commitment savings products can increase savings substantially. In another experiment in Uganda, Cadena and Schoar (2011) found that text message reminders increased loan repayment by nearly the same amount as sizable financial incentives.

In this paper, we explore if and how explicit promises can increase loan repayment in a real-world setting. Research in behavioral science suggests that explicit promises trigger intrinsic drivers of behavior, such as guilt aversion (Charness and Dufwenberg, 2006; Vanberg, 2008) and self-awareness (Duval and Silvia, 2002), which can be more influential than financial drivers. This paper contributes to existing literature in two ways. First, as far as the author can tell, there are no field experiments testing the impact of explicit promises on real-world decision making.[1] This paper represents a novel effort to do so. Second, most existing work focuses primarily on peer-to-peer promise scenarios, namely experimental games in which participants make promises to each other, rather than an institution, firm, or authority figure (Charness and Dufwenberg, 2006; Vanberg, 2008; Ellingsen and Johannesson, 2004).[2] Therefore, existing work may not generalize to contexts where legal and formal relationships bind economic actors, as they do here.

This paper presents a natural field experiment on the impact of explicit promises on loan repayment behavior, conducted with a partner firm making online loans. The firm operates in the market space typically associated with the "payday lending" industry, but seeks to distinguish itself by aligning incentives with the borrowers' financial interest to improve individual well-being. In the experiment, we use various forms of an explicit promise at loan initiation (in the form of an "honor pledge") to motivate loan repayment. From a policy perspective, studying the use of promises in the context of short-term loan arrangements can provide valuable insights into the behavior of financially-vulnerable populations. Social scientists have found that poverty-related concerns consume significant mental resources for the very poor, reducing their cognitive capacity and performance (Shah *et al.*, 2012). These findings suggest that financially-distressed households taking short-term loans are more susceptible to behavioral biases, which may result in decisions that are detrimental to their well-being. These biases can be especially damaging because most short-term borrowers have minimal access to other sources of credit, so firms are able to exploit these biases

[1]Though Shu *et al.* (2012) do conduct a field experiment involving signing an honesty pledge, they focus more on truthful reporting than on follow-up behavior.

[2]One exception is research on honor codes, mostly in the education literature.

without losing customers.

The experiment targeted 4,883 first-time borrowers with the firm. These borrowers were randomized into seven groups, using a 3x2 incomplete factorial study design. The control group received no honor pledge, and there were three honor pledge versions used as treatments: 1) signing a given honor pledge; 2) re-typing the same honor pledge as in treatment (1) before signing; or 3) coming up with a personal honor pledge to type and sign. To complete the 3x2 design, we also randomized whether or not borrowers who received an honor pledge treatment were reminded of the honor pledge they signed in the messaging sent prior to the repayment deadline. Using this design, we test the idea that making an explicit promise can have an effect on borrowers' behavior by giving individuals an intrinsic reason to repay, and how reminders about those explicit promises might influence behavior.

The results suggest that the honor pledges had a minimal impact on repayment and other related outcomes. Specifically, we tracked six outcome indicators in the aftermath of our experiment, intended to capture behavioral response to the treatments. Three of these variables showed consistent patterns of beneficial response to the various honor pledge treatments (repaying the loan; repaying the loan in exact accordance with the initial agreement—being a "perfect payer;" and repaying when signing up for a payment plan). However, none of the observed effects were statistically significant. Additionally, there is little evidence to suggest that the various versions of the honor pledge treatments had different impacts on the key outcome variables, nor any to support a positive impact of being reminded about the honor pledge in the days immediately preceding the loan repayment due date. In the paper, we consider a variety of reasons why this experiment may not have produced measurable impacts, including experimental design, sample size, a failure of lab results to generalize to a real-world setting, and alternative hypotheses and nuances around promises and their impact on behavior.

This paper proceeds as follows. Section 2.2 outlines existing literature and provides information on the policy context and motivation for this experiment. Section 2.3 outlines the experiment itself. Section 2.4 offers a brief outline of the empirical strategy for analyzing the data from the experiment. Section 2.5 presents results. Section 2.6 provides a discussion and briefly concludes.

2.2 Motivating Literature and Background

2.2.1 Promises and Honor Codes

Existing research on motivating behavior using explicit promise statements seeks to understand our intrinsic motivation not to violate pre-set agreements (Shu *et al.*, 2011; McCabe and Trevino, 1993). Most of these studies focus on honor codes in academic cheating contexts. Despite the existing research, however, the underlying reason for why these pledges influence our behavior is not well understood. We can think of the behavioral impact of explicit promises as operating through two main channels: 1) by priming a desire to live up to others' expectations (external motivation); and 2) by priming a desire to live up to our own expectations (internal motivation).

Experimental literature on "guilt aversion" suggests the power of promises to change behavior comes from external expectations (Charness and Dufwenberg, 2006; Vanberg, 2008). In Charness and Dufwenberg (2006), a person with guilt aversion is described as one who "suffers from guilt to the extent he believes he hurts others relative to what they believe they will get. Therefore, he is motivated by his beliefs about others' beliefs."[3] In Vanberg (2008), subjects who were given a chance to communicate prior to playing the classic dictator game would exchange promises, and as predicted by Charness and Dufwenberg (2006), these promises led to significant changes in participant behavior. In the experiment, 73% of dictators who made a promise to "Roll" (a move that was less beneficial for the dictator) stuck to their promises, whereas 52% of dictators who made no promises about their behavior chose to "Roll." Here, guilt aversion encourages behavior change by inducing psychological discomfort and pressure to conform to social expectations (Pelligra, 2011; Shu *et al.*, 2012).

[3]Charness and Dufwenberg (2006, p. 1583)

Similarly, in a series of experiments, Shu *et al.* (2012) find that dishonest behavior was reduced when participants signed an agreement before facing an opportunity to cheat. For example, in a field experiment with an insurance company testing the effect of signing an honesty pledge on an automobile policy review form before reporting their current odometer mileage ("I promise that the information I am providing is true"), customers who signed prior to completing the form reported significantly higher mileage than those who signed at the end. In other words, signing prior to reporting decreased the rate of customer cheating. The authors also used a word-completion task to measure how signing an honesty pledge before facing an opportunity to cheat on tax return forms changes the accessibility of words related to ethics and morality. Specifically, they found that those who signed the pledge prior to filing generated more ethics-related words than those signing at the bottom of the form. The authors posit that signing might reduce dishonest behavior by increasing the saliency of moral standards.

Self-awareness theory, on the other hand, suggests that honor codes are powerful because of the internal motivators they trigger. According to this theory, explicit promises direct individuals' attention to their decisions, which can elicit reflection and behavior change (Duval and Silvia, 2002). Additionally, identity theory suggests that this promise effect is especially powerful if individuals feel their promises relate closely to their self-identities, since one's self-identity creates a set of expectations that guide behavior (Stets and Burke, 2000).

Research in effort justification further suggests that explicit promises may be especially effectual when individuals identify personally with their commitment. For example, in a study on the value of self-made products, Norton *et al.* (2011) found using a set of experiments that creating an object increases its value to its creators—a phenomenon the authors call the "IKEA effect." One pair of experiments established that participants were willing to pay more for self-assembled IKEA boxes and their own origami creations than they were willing to pay for identical pre-assembled IKEA boxes or origami creations made by experts. They also found that the increased value of the self-made object was a result of the creator's ability to identify with their creation rather than a result of the endowment effect, the length of ownership, or physical contact with the objects.[4] Based on this study, we might expect that triggering self-identity may increase a promise's value to individuals, activating an stronger intrinsic motivation to follow through.

Regardless of the distinction between external and internal motivation, research on honor codes overall suggests that using seemingly trivial manipulations, such as moving an honesty pledge and the signature on a self-report, can greatly alter the power of promises (Shu *et al.*, 2012). This could be because such manipulations direct an individuals' attention to their behavior in a different way, increasing the emotional salience of the promise (Shu *et al.*, 2012). Therefore, we test the hypothesis that salience is important in the context of loan repayment, and seek to determine whether the effectiveness of an honor code relies on how borrowers are presented with, and remember, the honor code.

Finally, it is important to distinguish between peer-to-peer promises, which dominate current experimental evidence on the effect of promises on behavior, and promises between clients and firms. For example, Charness and Dufwenberg (2006) and Vanberg (2008) paired participants up in dictator games, while studies on the effect of computer-mediated communication on cooperation in social and prisoner's dilemmas has focused on interaction between equals.[5] However, some research on explicit promises has extended past these peer-to-peer interactions. Notably, work on academic honor codes focuses on promises by students to educational institutions, broadly speaking (McCabe and Trevino, 1993; Mazar *et al.*, 2008), while one of the experiments in Shu *et al.* (2012) targets a car insurance company and its clients. In this paper, we contribute by exploring promise effects in the client-firm context using a field experiment, focusing on how promises can influence financial behavior between borrowers and lenders—with money at stake.

---

[4]For research on individuals' preference for endowed goods, see Kahneman *et al.* (1990). For research suggesting that time spent in physical contact with objects can increase feelings of ownership and value, see Peck and Childers (2003) and Peck and Shu (2009).

[5]For a review of such studies see Bicchieri and Lev-On (2007).

2.2.2 Payday Lending and Borrower Biases

In the United States, there are roughly twice as many payday lending locations as there are Starbucks Coffee locations.[6] These locations offer short-term, high interest rate loans to individuals who need money from their next paycheck in advance. According to the Pew Charitable Trust, roughly 12 million Americans take out payday loans each year, and a disproportionate number of these individuals are low-income (Urahn, 2011). Roughly 80% say they would need to cut back on expenses like food and clothing without these loans, suggesting that most use these loans for basic necessities, not unexpected emergencies (Urahn, 2011; O'Neil, 2011). Faced with mounting bills and limited access to other sources of credit, many borrowers say payday loans are "a good solution" for their financial issues.[7]

However, payday loans are heavily criticized by the media and federal regulators because of their high cost to borrowers and the feeling that they are exploitative financial products.[8] Most payday borrowers have limited access to credit, and are reliant on these loans despite their exorbitant interest rates and fees. As a result, many borrowers are forced to take out multiple loans to repay outstanding debts, initiating a debt spiral. According to the Consumer Financial Protection Bureau, roughly 50% of loans are made to borrowers trapped in a debt spiral of 10 or more loans.[9]

In an effort to lessen the harmful impact of payday loans, many states have passed bills limiting the number of loans an individual can take out in a given year, and the Consumer Financial Protection Bureau is working on a new set of rules to govern the industry nationally as of February 2015.[10] Similarly, the UK's Financial Conduct Authority finalized new rules for payday lenders in late 2014, limiting the cost of loans by capping the

[6] CFRL (2014)

[7] Bazelon (2014)

[8] Gensler (2014); Oliver (2014)

[9] Cordray (2014)

[10] Kiel (2013a); Silver-Greenberg (2015)

daily fees and interest rates allowable for such loans.[11] These laws have a mixed record. In some cases, lenders exploit loopholes in the law to continue lending, while in other cases consumers cross state borders themselves to obtain loans.[12] Furthermore, there may be unintended consequences associated with successfully limiting a borrower's ability to take out these loans. One potential borrower, Evelyn Reese, a 70-year-old social security recipient, suggests that such an outcome would be "a terrible mess for people who live from week to week."[13]

While critics describe payday loans as predatory, lenders argue that a simple economic model of risk and return justifies the interest rates they charge. Specifically, they point to a high default risk on these loans—a study by the Consumer Financial Protection Bureau, which analyzed data from payday lenders over an 11-month period in 2011 and 2012, found that approximately 20% of payday borrowers defaulted at least once during that period.[14] In an economic model with rational, optimizing borrowers, and holding all else constant, high interest rates on payday loans should keep individuals with no ability to repay from borrowing. However, these loan characteristics do not seem to impact consumers' borrowing decisions. According to Shah *et al.* (2012), this is due to the impact of resource scarcity on borrowers' decision-making process. When a borrower's attention is consumed by the possibility of not being able to afford ordinary living expenses, present-bias might drive them to neglect the deferred costs of the loan—including its high interest rates. This could explain financially-vulnerable individuals' tendency to borrow and suggests a solution to a problematic element of payday loans: shifting borrowers' attention to their loans and consequent commitments.

As illustrated by the current situation with payday loans, research in social science has consistently found that financial incentives and extrinsic motivation, like high interest rates

[11]FCA (2014)

[12]Kiel (2013b)

[13]Rivlin (2011)

[14]CFPB (2014, p. 26). The default rate in the data for this experiment was 20.60%.

and limits, can fail to influence behavior as theory might predict. The high interest rates on payday loans do not motivate borrowers to avoid loans they are unable to repay, nor do they seem to prevent borrowers from taking out additional loans and trapping themselves in a debt spiral. As work on attentional neglect argues, this behavior may persist in part because individuals are not paying attention to the financial incentives. This suggest that while payday borrowers' inability to repay existing loans is primarily the result of limited resources, the consistent failure to repay may also be related to behavioral factors: repayment may not be salient when payment is due, borrowers may struggle to follow through on their commitment to repay, or present-bias might lead borrowers to incorrectly discount future costs and benefits. The treatments in this experiment are motivated by the idea that intrinsic motivators may help promote repayment more than financial ones, due to their ability to shift attention and engage individuals in proactive behaviors.

However, using behavioral motivators to influence financial decisions is not simple, and results from past experiments in this area are mixed. Experiments on loan repayment using peer pressure (Breza, 2010) and reminders (Cadena and Schoar, 2011) illustrate both their effectiveness and nuances. For example, Breza (2010) presents a field experiment on the effect of peer pressure on repayment rates, finding that "repayment peer effects can create both positive and negative incentives for borrowers."[15] An experiment by Karlan *et al.* (2015) found that reminders for loan repayment improve repayment only when the reminder included the loan officer's name—an effect that did not hold for first-time borrowers.

This paper tests the effect of explicit promises on decision-making processes using a randomized, natural field experiment, a methodology uncommon in the extant literature in this area. The experimental design allows us to tease out potential effects and better understand both if promises influence behavior, and if they do, why they do. Further, we hope that the nature of the relationship captured in our experiment—a lender and its borrowers—will contribute to the literature by extending existing work into a real-world scenario with money at stake.

[15] Breza (2010, p. 1)

2.3 Experiment Overview

2.3.1 Motivating Literature for Experimental Design

The designs of the experimental treatments in this study are motivated by four branches of existing behavioral literature; namely, research on the effects of explicit promises, salience, the "IKEA effect," and reminders. This literature builds on psychological theories of morality, guilt aversion, and self-awareness. We outline this motivating literature and relevant hypotheses that stem from the literature here.

Explicit Promises

Generally, payday lenders do not require that borrowers make an explicit promise to repay. However, research suggests that signing an honor pledge may trigger behavior change because of our desire to perceive ourselves as moral (Aquino and II, 2002). If we fail to adhere to the pledge, it could lead to psychological discomfort, commonly known as cognitive dissonance.[16] The potential for discomfort motivates us to change our behavior and adhere to prior statements and commitments. This underlies all experimental treatments and leads to a simple hypothesis, namely that any form of honor pledge will increase repayment.

Salience

Existing research on honor pledges suggests that the pledges need to be particularly salient to trigger cognitive dissonance (Shu *et al.*, 2012). Research on the impact of directed attention and salience suggests that simple manipulations of individuals' level of attention can determine specific behavior (Jonas *et al.*, 2002; Duval and Silvia, 2002; Mazar *et al.*, 2008; Shu *et al.*, 2012). Much of this research is guided by Duval and Wicklund (1972), which presents a theory of objective self-awareness. This theory posits that forcing individuals

[16] See Stone and Cooper (2001) for a review of three major theories of cognitive dissonance and a proposed fourth one, all of which support the use of honor codes to motivate behavior change, albeit through different mechanisms.

to pay attention to themselves and their behaviors enables self-evaluation and behavior change. Thus, we can hypothesize that the more salient an honor pledge, the more likely it is to change behavior. Note that we refer here to the level of salience at the moment of the pledge, not immediately in advance of repayment (which is discussed later in this section). We test this experimentally using treatments that vary the salience of honor pledges at loan initiation, with the hypothesis that more salient honor codes will be more effective at increasing repayment.

The IKEA Effect

The effectiveness of an honor pledge may also rely on the value borrowers place on the pledge and how they relate to it (Stets and Burke, 2000; Turner, 2012). Making the pledge closely relate to a borrower's self-identity, for example, could increase the likelihood of engagement with the content of the statement (Kettle and Habul, 2011). In Norton *et al.* (2011), the researchers note that increased customization and perceived control over self-made products can make them more valuable to their creators. While their study relates specifically to physical objects, literature on effort justification has demonstrated that effort can increase the value of non-physical objects as well, like participation in a discussion group (Aronson and Mills, 1959). Consequently, we vary the extent to which an honor code is personalized, to test the hypothesis that personalized honor codes are more likely to be adhered to than standardized ones.

Reminders

Finally, behavioral researchers argue that reminders about repayment can on their own increase repayment rates (Cadena and Schoar, 2011; Karlan *et al.*, 2015). For example, Cadena and Schoar (2011) posit that poor repayment behavior is a result of limited financial planning, and suggest that text message reminders increase repayment rates by keeping debt salient, which helps borrowers better manage their finances. In this experiment, all participants receive email reminders about repayment, but half receive a reminder that includes an additional reminder about the honor pledge made at loan initiation. This makes the honor pledge differentially salient at the moment of repayment across treatment groups, and enables us to test the hypothesis that salience is important not only at the moment of a pledge, but also when an individual is asked to follow through on their commitment.

2.3.2 Experimental Design

Partners

My research partner for this field experiment was a startup firm based in San Francisco, CA, which makes short-term loans to qualified borrowers through an online platform. Their mission is to provide short term loans in a non-exploitative manner, at the lowest rates possible given the default risk inherent in these loans. Their primary method of customer engagement is through their website, where users apply for loans, get approval, and have money sent to their bank accounts in a very short window of time.

Subjects

Subjects in this study consisted of 4,883 recipients of loans through the firm over a six-month period, from March 2013 to September 2013. The subjects were not recruited—they were customers who came to the website freely to take a loan, and consented to communication through the firm as part of the borrowing agreement. No inducement was offered for their participation, and they were unaware of the experimental nature of the borrowing process, as described here. All borrowers were first-time borrowers, meaning they had never obtained a loan from the site previously.

Study Design

Upon initiating a loan process through the firm's online platform, subjects were immediately and automatically randomly assigned into one of seven different groups—one control group and six treatment groups. All groups had a similar online process to obtain the loan, consisting of seven screens before the approval confirmation screen. Screenshots for the onboarding loan process from the perspective of the borrower are included in Appendix B.1. Importantly, the first six screens were the same for all individuals regardless of treatment/control group. Only the seventh screen that subjects saw was different for the control and various treatment groups. This ensured to the greatest extent possible that differential attrition due to the treatment was minimized (as sharing of personal information, a potential hurdle for skeptical borrowers, was done on pre-treatment screens).

Treatments and Controls

The experiment uses a 3x2 incomplete factorial study design, with a control group not receiving an honor pledge message at any stage. The 3x2 setup for the six experimental treatments operated as follows. First, subjects were randomly assigned to receive one of three honor pledge "treatments" on the final pre-approval screen during the loan initiation process—referred to hereafter as the "Simple," "Copy," and "WriteIn" treatments. These three treatments offered different versions of an "honor pledge" for the subjects to interact with, specifically: 1) the "Simple" treatment required subjects to read and sign a given, default honor pledge; 2) the "Copy" treatment required subjects to re-type the same given, default honor pledge word-for-word into a text box, and sign it before proceeding; and 3) the "WriteIn" treatment required subjects to come up with their own honor pledge, write it into a text box, and sign it before proceeding. Note that in all cases, the font used for the signature was "handwriting" font to increase realism. Figure 2.1 shows the differences between the control and three treatment screens (the full screens seen by subjects in the different treatments are provided in Appendix B.2).

Second, the subjects also differed in the type of repayment reminder they received, three days before the repayment deadline. Specifically, subjects were randomly assigned to receive either the "standard" reminder email, or a reminder that made explicit reference to the honor pledge the subject signed. Note that all subjects, including the control subjects, received the reminder email. Henceforth, the treatments that included a reminder of the honor pledge will be labeled with "HP" (so, "Simple-HP," "Copy-HP," and "WriteIn-HP").

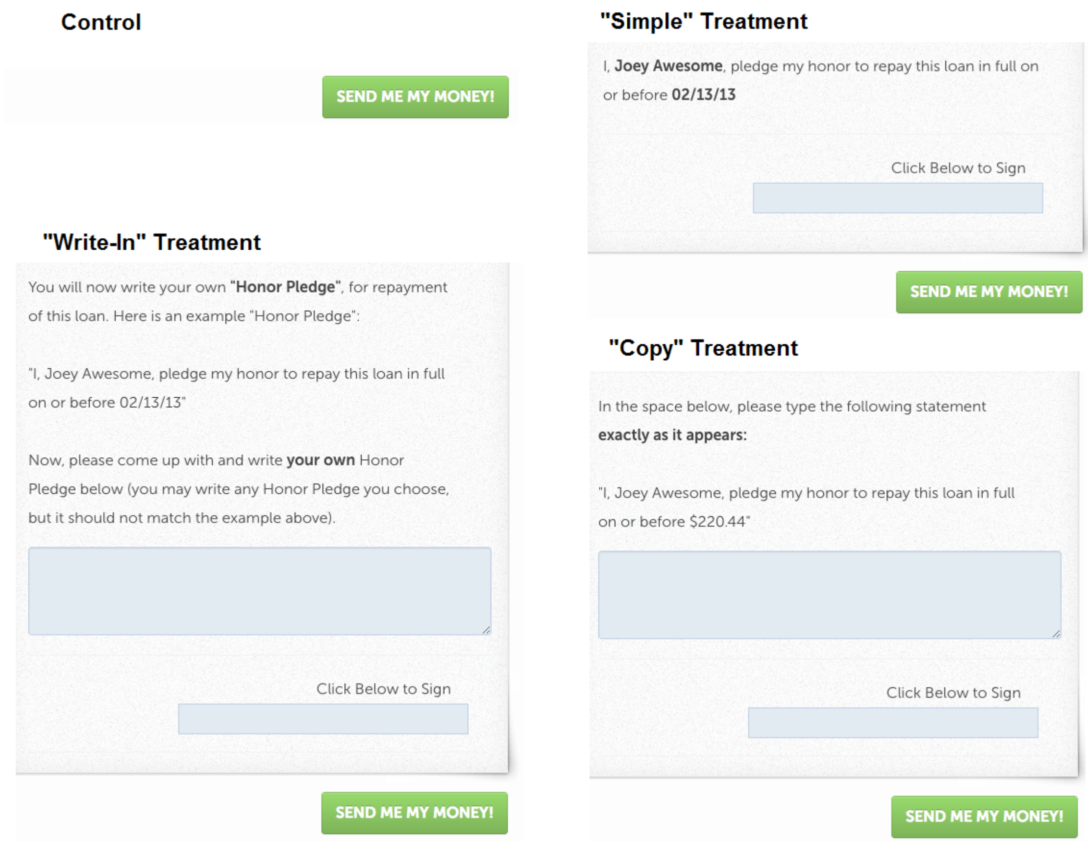


Figure 2.1: *Final Pre-Approval Screens for Control and Three Treatments*

Table 2.1 summarizes the differences in the control and treatment groups, for clarity.

Fees and Logistics of the Loan Process

Borrowers took out loans for between 7 and 30 days through the site. Loans ranged from \$100-\$250, with a mean loan size of \$219.10. The vast majority of loans were for \$200 (36.3%) or \$250 (50.9%). If the chosen loan duration at initiation was under 30 days, subjects were allowed to extend their loans up to the 30 day maximum at any time, without penalty. The fee for the loan was 17.6% of the amount of money borrowed, with a 30 cent-per-day reduction in fee for early repayment. Individuals who were late in repaying their loans could opt to enter into a payment plan with the firm, negotiated on a case-by-case basis. There was a flat \$15 fee for late repayment. This late repayment fee was reduced by 30

Table 2.1: *Control vs. Treatments (3x2 Incomplete Factorial Design)*

	Control	**Simple**	**Simple-HP**	**Copy**	**Copy-HP**	**WriteIn**	**WriteIn-HP**
Honor Pledge	None	Simple	Simple	Copy	Copy	WriteIn	WriteIn
Reminder	Standard	Standard	w/Pledge	Standard	w/Pledge	Standard	w/Pledge

cents per day if the initial loan was for under 30 days, in line with the allowance of loan extensions and the early repayment fee reduction.

Importantly, borrowers were set up for auto-repayment of their loans through an ACH transfer upon loan initiation. In other words, the "default" was for the loan repayment to be initiated at the time agreed upon for repayment at the initial loan agreement. This is important because it may lead us to underestimate the behavioral effects of this intervention—since the default is to repay, it may be the case that some individuals repay because it is the default, and not because they make an active choice to repay.

2.3.3 Data and Baseline Characteristics

The data was collected by the firm, and not the researcher. The firm also collected data on repayment status, loan amounts and dates of agreement, and demographic information on the borrowers, as required by law. The specific data that the firm collected from each borrower are: name, address, phone number, social security number, employer, bank, channel data (IP address, browser, etc), and behavioral data (user actions on firm's website, like logins, page clicks, etc.). All the data the firm gathered was with the consent of the customer, as required by the Fair Credit Reporting Act for lenders. The firm scrubbed the data of all identifiers, removing data on name, address (except zip code), phone number, social security number, and any other data that could be traced back to an individual, including IP address data. The firm then replaced this data with "borrower ID numbers," added a treatment/control group variable, and shared the data with the researcher.

Descriptive Data and Baseline Characteristics

We observe data from 4,883 first-time borrowers. Appendix Table B.1 outlines the number of subjects in each treatment and control group, and summarizes the demographic variables available for individuals in these groups. In addition, Appendix B.3 provides visuals depicting the distribution of subjects' job status, age, and income. Note that the number of observed demographic variables is limited, due to legal restrictions around sensitive financial transactions.

Two particular elements of the demographic data are worth mentioning, for clarity. First, subject gender was not in the data, but the partner firm was able to generate a variable ("Probability Male") that captures the probability that a given individual was male or female using an internal algorithm based on the individual's first name. The mean and standard deviation of this variable are reported in Appendix Table B.1, along with a dummy variable for gender, created using the "Probability Male" variable (which considers a person male/female if the algorithm judged them to be male/female with probability 99% or more). Second, the "Income" variable in the data is self-reported by each subject, and has large outliers (with nine observations with self-reported income greater than $1,000,000, as visible in Appendix B.3, Panel (B.3.3)). This largely explains observed variation in mean income across treatment groups.[17]

Randomization Check

While some recent papers have questioned the need for randomization checks in experiments (Mutz and Pemantle, 2011), we argue that randomization checks are necessary in this analysis for two main reasons. First, the randomization process was conducted by the firm and not the researcher. Though the firm has a track record of experimentation and a background in randomization procedures, the randomization check remains important.

[17] Notably, the removal of observations where income exceeds $1,000,000 reduces the coefficient on the "Copy" treatment in the randomization check in Appendix Table B.2, column (6), from $20,280.1 to an insignificant $1,668.34.

Second, there is the potential for attrition during the loan onboarding process. Appendix Table B.1 shows that there is imperfect balance in terms of the number of subjects in each treatment group. This is due to the fact that subjects were randomly assigned at the initiation of the sign-up process. Therefore, any attrition during the multi-screen sign-up process would cause an imbalance. Importantly, attrition on the first six screens would not have been caused by the differences in treatment, which only happened on the seventh screen. However, it is possible that any one of the treatments increased attrition on this seventh screen. Though we cannot know this for sure, the fact that the control screen had the highest number of subjects in the final data suggests that it is possible that there was differential attrition. Therefore a randomization check is essential, to ensure balance amongst those who completed the sign-in process.

Appendix Table B.2 presents the results of the randomization check. Specifically, the table presents the coefficients of regressions of the various demographic characteristics (shown as y_i below) on dummy variables for the six treatment groups (shown as T_k, with k ranging from 1-6 for the six treatment groups, below), omitting the control group. The econometric model is as follows:

$$y_i = \beta_0 + \sum_{k=1}^{6} \beta_k (T_k)_i + \varepsilon \tag{2.1}$$

A series of f-tests were conducted to evaluate the joint-significance of the coefficients for each demographic variable regression. The f-statistics and p-values are reported in Appendix Table B.2. We find no joint significance for the dummy variables associated with the treatment groups, suggesting no strong imbalances between control and treatment groups from potential attrition issues. Nevertheless, we control for demographic characteristics in the analysis.

Outcome Variables

We tracked a number of repayment-relevant variables in the data over time amongst experimental subjects. In particular, we focused on the following six outcome variables: 1)

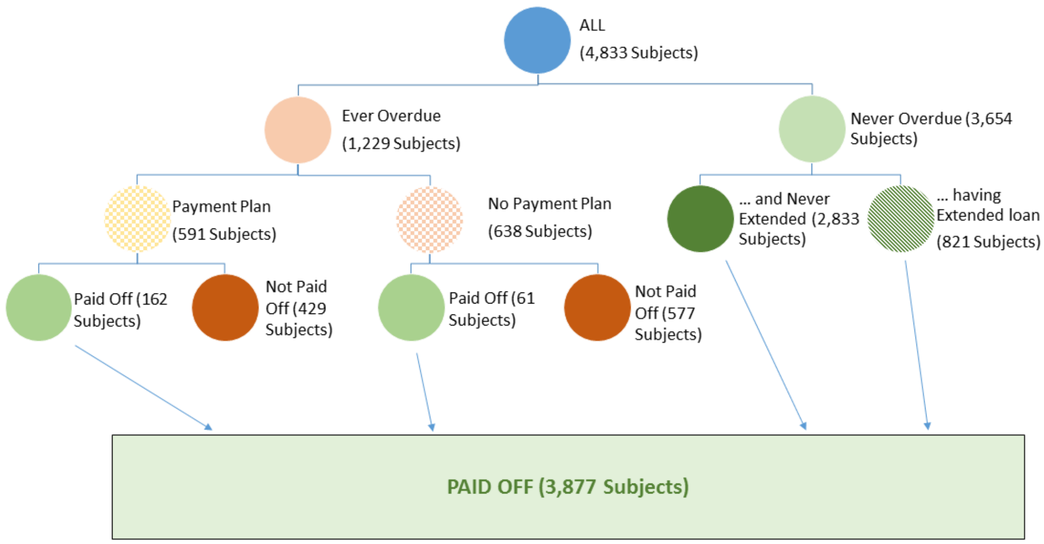


Figure 2.2: *Repayment Process and Related Variables*

whether the subject ever paid off their loan; 2) the number of days between loan initiation and loan repayment (for those who repaid only); 3) whether the subject was a "perfect payer," meaning they repaid the loan in full by the original due date they selected when obtaining the loan; 4) whether the subject was ever overdue in repaying their loan; 5) whether an overdue subject chose to go on a payment plan or not; and 6) whether an overdue subject on a payment plan actually repaid their loan. To understand why these specific variables were selected for analysis, it is important to understand how the repayment process and the various measured variables relate to each other. Figure 2.2 provides a visual to show how borrowers moved through the repayment process, and how those movements relate to the variables we use as behavioral outcomes.

As visible in Figure 2.2, individuals who repaid their loan could have done so in four different ways: 1) by repaying in accordance with their initially selected due date (being a "perfect payer"); 2) by extending their initial repayment date and repaying in full by that new due date (recall that individuals who chose loan periods of under 30 days could extend to 30 days without penalty); 3) failing to repay by their due date, setting up a payment plan directly with the firm, and repaying via this payment plan; or 4) failing to repay by

Table 2.2: *Outcome Variables*

Outcome Variable	Relationship to Promises and Treatments	Abbreviation
1) Subject paid off loan (dummy)	Likelihood of being paid off in full is a big-picture measure of the efficacy of the explicit promise treatments relative to the control.	Paid Off
2) Subject was a "perfect payer" (dummy)	Likelihood of sticking to the agreed due date is a measure of the power of the explicit promise to repay in the treatments, relative to the control.	Perfect Payer
3) Days between loan initiation and repayment	If the mean days to repay is smaller for a given treatment, this signifies a greater aggregate willingness to repay the loan caused by treatment.	Days to Pay
4) Subject was overdue (dummy)	Being overdue is a sign of a failure to stick to the loan agreement at initiation, thus a failure to follow through on the commitment made.	Ever Overdue
5) Overdue subject chose to go on a pay plan (dummy)	Going on a payment plan when overdue is a sign that a subject wants to repay despite failing to do so by the due date. This variable can capture a commitment to repay despite initial difficulty.	PayPlan
6) Overdue subject on a pay plan repaid (dummy)	Repaying via a payment plan captures a commitment by a delinquent borrower to repay the loan, and a willingness to work to do so. This variable is therefore a measure of genuine commitment to the initial agreement to repay.	PayPlan Paid

their due date, but repaying eventually through a direct payment to the firm. The outcome variables we select for analysis here are based on these channels for repayment and how the use/non-use of these payment options might relate to the explicit promise subjects made at loan initiation. Table 2.2 outlines the relationship between these six outcome variables and a justification for their connection to the explicit promise treatments.

Note that while we track six outcome variables, there are differences in how important these variables are as indicators of behavioral response. This is especially true since some of these variables are conditional on previous behavior or harder to interpret cleanly. In particular, note that outcome variables (5) and (6) in Table 2.2 are conditional on being overdue to begin with, a state that is itself a potential function of treatment. Additionally, variable (3) in Table 2.2, days until payment, is conditional on having repaid the loan at all. Because of this, it is not obvious whether a high or low value for this variable is a good thing—on the one hand, taking longer to repay is a sign of a failure to follow through on your commitment, but it is better to take a long time to repay than not to repay at all. Indeed, given how these variables interact, we can think of there being two "main" variables that best capture behavioral response—variables (1) and (2) (paying off the loan, and being a "perfect payer," who repays by the date initially agreed). Therefore, when we assess the results, we will pay particular attention to these two variables.

2.4 Empirical Strategy

2.4.1 Average Treatment Effects

Analytically, we first compute average treatment effects on the outcomes outlined using regressions, with a linear probability model specification for instances where the dependent variable is binary. The regression specification is as follows:

$$y_i = \beta_0 + \sum_{m=1}^{6} \beta_m (T_m)_i + \gamma_i + \eta_i + \varepsilon \tag{2.2}$$

We control for loan characteristics (loan amount and days of the loan, shown as γ_i) and borrower demographics (job type, pay frequency, pay type, log of income, and age, shown as η_i) in these regressions. Note that we do not control for gender, since we do not have a certain measure of borrower gender (though we do incorporate the gender variable in section 2.4.3).

2.4.2 Grouping Treatments and Isolating Reminder Effects

One plausible ex-ante hypothesis might be that the addition of a "reminder" of the honor pledge might have a minimal impact on the outcome variables, particularly given that all subjects received some form of a reminder email from the firm. For example, if people did not read the reminder email carefully or simply deleted it without reading it, we would expect no significant differences between subjects receiving a given honor pledge version with or without the added reminder. If that is the case, an alternative specification that bundles honor pledge treatments and controls for a common "reminder effect" for the treatments would be a plausible way to estimate treatment effects as well. Indeed, this specification would have the added benefit of increasing the sample size within each treatment and reducing standard errors.

The process to generate a specification for this analysis is as follows. First, we create three dummy variables for "All Simple," "All Copy," and "All WriteIn" treatments, which combine subjects in each honor pledge group that did/did not receive the added honor pledge reminder (these dummy variables are shown below as ($\sum_{m=1}^{3}(TG_m)_i$). We then create a dummy variable (shown as $PledgeTreat * Reminder$ below) that interacts a dummy for being in any honor pledge treatment at loan initiation (in other words, being in any group except the control—shown as *PledgeTreat* below) with a dummy for receiving a reminder email that makes explicit reference to the honor pledge (shown as *Reminder* below). We then run the following regression specification below, with added controls for loan characteristics (loan amount and days of the loan, shown as γ_i) and borrower demographics (job type, pay frequency, pay type, and log of income, shown as η_i):

$$y_i = \beta_0 + \sum_{m=1}^{3} \beta_m (TG_m)_i + \beta_4 (PledgeTreat * Reminder) + \gamma_i + \eta_i + \varepsilon \tag{2.3}$$

The β_1, β_2, and β_3 regression coefficients in this specification estimate the average treatment effect of each honor pledge version (Simple, Copy, and WriteIn) without the honor pledge reminder, while the β_4 coefficient estimates the effect of the honor pledge reminder across all honor pledge treatments. Note that we also run an alternative version of this specification that omits the $\beta_4(PledgeTreat * Reminder)$ term, and simply estimates the average treatment effect of the three honor pledge versions used without isolating a unique and constant "reminder effect" across treatments.

2.4.3 Disaggregated and Conditional Average Treatment Effects

Finally, we compute disaggregated and conditional average treatment effects based on prior characteristics. In particular, we focus on three covariates: income, age, and gender. We select these covariates because there are plausible reasons for why explicit promises might be differentially impactful within these covariate groups. As we have discussed, one might reasonably suspect that low-income borrowers are more prone to behavioral biases around loan repayment, and therefore more likely to be influenced by the behavioral treatments than high-income borrowers.

Similarly, we explore the idea that being forced to make an explicit promise and a commitment to repay might have different impacts based on borrower age. There are multiple hypotheses possible here. We might posit that younger borrowers are less likely to be self aware about their financial decisions and their consequences than their adult counterparts, making an explicit promise more likely to influence their behavior. Alternatively, one might suspect that adults are more likely to take a promise seriously than a younger borrower.[18] We hope our analysis can provide some suggestive evidence.

Finally, there is a large body of literature in psychology and economics looking into how decisions differ by gender,[19] and this experiment provides an opportunity to gain evidence on how explicit promises might have differential impacts by gender. Ex-ante, we suspect that women are more likely to respond to experimental treatments that men, drawing on findings in existing research that women may be more likely to think of communal wellbeing (Eagly and Wood, 1999, p. 413) or contextual factors (Eckel and Grossman, 1996, p. 154) when making decisions, two features we might associate with the explicit promise treatments.

[18] In terms of self-awareness theory, there is reason to suspect that self-awareness varies a great deal within individual adults. As Rochat (2003) notes: "As adults, we are constantly oscillating in our levels of awareness: from dreaming or losing awareness about ourselves during sleep, to being highly self-conscious in public circumstances or in a state of confusion and dissociation as we immerse ourselves in movies or novels."(Rochat, 2003, p. 728)

[19] Eagly and Wood (1999) and Croson and Gneezy (2009) provide overviews of this work in psychology and economics, respectively.

Empirical Approaches

For the income analysis, we break subjects into quintiles based on self-reported income, and estimate conditional average treatment effects by quintile using the same specification as in section 2.4.1.

For age, we interact a continuous variable for subject age with the treatment variables to estimate how the treatment effects may have differed by age. The regression specification for this analysis is as follows, with controls for loan characteristics (loan amount and days of the loan, shown as γ_i) and borrower demographics (job type, pay frequency, pay type, and log of income, shown as η_i):

$$y_i = \beta_0 + \sum_{m=1}^{6} \beta_m (T_m)_i * Age + \sum_{m=7}^{12} \beta_m (T_m)_i + \beta_{13}(Age) + \gamma_i + \eta_i + \varepsilon \tag{2.4}$$

The $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$, and β_6 regression coefficients are the key coefficients in this specification, as they provide us with an estimate of how the outcome variables differ by age across the treatments.

Finally, for the gender analysis, we first restrict attention to subjects whom the gender algorithm (described in section 2.3.3) reports as being either male or female with at least 99% certainty.[20] We then interact a dummy variable for "male" with the treatment variables, to obtain estimates of the differences in treatment effects by gender. The regression specification for this analysis is as follows, with controls for loan characteristics (loan amount and days of the loan, shown as γ_i) and borrower demographics (job type, pay frequency, pay type, age, and log of income) shown as η_i:

$$y_i = \beta_0 + \sum_{m=1}^{6} \beta_m (T_m)_i * Male + \sum_{m=7}^{12} \beta_m (T_m)_i + \beta_{13} Male + \gamma_i + \eta_i + \varepsilon \tag{2.5}$$

Again, the $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$, and β_6 regression coefficients are the key coefficients, providing an estimate of the differential impact of the various treatments for men relative to

[20] Individuals with "male probability" of 99% or more were classified as male, while those with "male probability" of 1% or less were classified as female. All others, including those with first names missing, were labeled as ambiguous. This resulted in 1358 "males," 1775 "females," and 1750 "ambiguous" cases.

women. That said, the $\beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}$, and β_{12} coefficients are also important, as they provide an estimate of average treatment effects for women only, across treatment versions.

2.5 Results

Appendix Tables B.3-B.12 present a comprehensive report of the regression results, reporting the estimated treatment effects for each outcome variable. Recall again that while we report all six outcome variables measured, we are especially interested in two variables: "Paid Off" and "Perfect Payer." Results at the aggregate level (section 2.4.1) are presented in Appendix Table B.3, with controls, while Appendix Tables B.4 and B.5 report grouped and honor pledge reminder isolating estimates at the aggregate level (section 2.4.2). Appendix Tables B.6-B.12 report subgroup estimates (section 2.4.3). Though all estimates reported include demographic control variables, regressions without controls provide similar results and will be provided in the online appendix for this paper.

The aggregate analysis (Appendix Table B.3) suggests that the honor pledge treatments had small positive effects on three of the outcome variables (Paid Off, Perfect Payer, and PayPlan Paid) and mixed effects on the other three variables. However, none of these effects are statistically significant, suggesting that the treatments had a minimal impact on aggregate outcomes. In Table 2.3, we present the expected values of the six key outcome variables for the different treatment groups, assuming "average" population values for the covariates used as controls in the regressions, to provide a sense of effect size. Table 2.3 also reports t-statistics for the differences between control group and the individual treatment groups in parentheses.

Notably, subjects in all honor pledge treatments were more likely to pay of their loans, be a "perfect payer," and pay off given signup to a payment plan than in the control group. Note that the former two variables (Paid Off and Perfect Payer) are those that we consider the best indicators of behavioral response. The effect sizes for these two variables range from roughly 0-3%. However, none of these effects were statistically significant, and joint hypothesis tests on all treatment coefficients confirm that we cannot reject a null hypothesis

Table 2.3: *Estimated Outcome Values for "Average" Subjects, by Treatment (t-statistics in parentheses)*

Outcomes	**Control**	**Simple**	**Simple-HP**	**Copy**	**Copy-HP**	**WriteIn**	**WriteIn-HP**
1) Subject paid off loan	78.3%	80.3% t=0.90	78.7% t=0.18	80.8% t=1.18	79.5% t=0.58	79.1% t=0.36	79.0% t=0.33
2) Subject was a "perfect payer"	56.8%	57.4% t=0.21	57.5% t=0.25	59.7% t=1.13	58.7% t=0.75	58.8% t=0.78	57.2% t=0.15
3) Days until loan repayment	26.2	26.9 t=0.80	26.5 t=0.38	26.5 t=0.34	26.1 t=-0.14	26.1 t=0.08	26.9 t=0.72
4) Subject was overdue	25.2%	24.8% t=-0.16	26.4% t=0.52	23.8% t=-0.61	25.4% t=0.08	25.1% t=-0.03	25.5% t=0.12
5) Overdue subject chose a pay plan	48.8%	50.7% t=0.37	51.1% t=0.45	48.0% t=-0.16	47.9% t=-0.16	47.3% t=-0.29	43.0% t=-1.15
6) Overdue subject on pay plan repaid	22.3%	33.0% t=1.60	27.9% t=0.91	27.0% t=0.71	25.9% t=0.55	27.9% t=0.84	28.2% t=0.88

that the various honor pledges had no effect on the key outcome variables. Additionally, there is no clearly discernable pattern to these effects across honor pledge versions—there is little to suggest that any individual honor pledge treatment performed differentially better on aggregate.

Appendix Table B.4 groups the six treatments into three broader treatments, by honor pledge version (by ignoring honor pledge reminders), and reports the results from regressions of the key outcome variables on these three treatment categories. Again, while this table does offer some suggestive evidence for positive treatment effects of the honor pledges on the likelihood to paying off loans, being a "perfect payer," and paying off given signup to a pay plan, the effects are not statistically significant.

Appendix Table B.5 enables us to isolate the unique effect of a reminder of the honor pledge on the key outcome variables empirically, by replicating the specification in Appendix Table B.4, with the added interaction term as described in section 2.4.2. Across the board, reminders of the honor pledge had no clear effect on outcomes; indeed, the signs on the honor pledge reminder coefficients suggest a small negative effect of reminders on all six outcome variables. This is a notable result—not only did reminders of the honor pledge

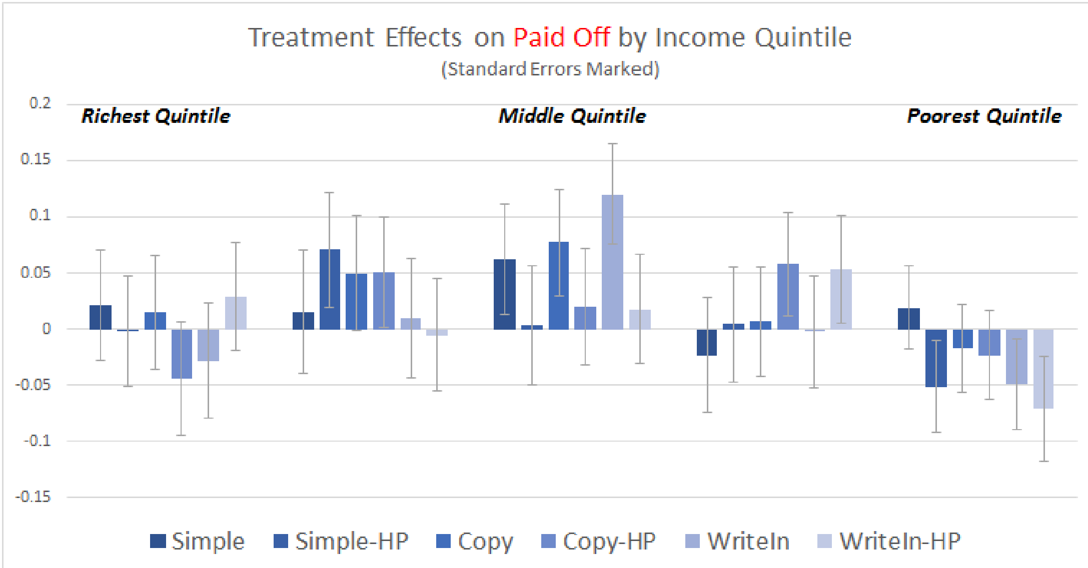


Figure 2.3: *Average Treatment Effects on Paying Off Loan, by Income Quintile*

signed at loan initiation not increase repayment, there is more suggestive evidence that they fractionally hurt key outcome behaviors than anything else.

Overall, the analysis at the aggregate level suggests that the treatments had a small impact on some of the outcome variables of interest, and in particular on the two variables we care most about—paying off a loan and being a perfect payer. One possibility is that a lack of significant effects at the aggregate level might mask heterogeneities of effects among subgroups. We explore this possibility using the specifications outlined in section 2.4.3, beginning with disaggregated analysis using income quintiles. Appendix Tables B.6-B.10 show regressions providing estimates of average treatment effects for each of the income quintiles in the study on all six key outcomes. In Figures 2.3 and 2.4, we plot the regression coefficients estimating the average treatment effects for the two main outcome variables—the probability of paying off a loan and being a perfect payer, by income quintile (specifications (1) and (2) in Appendix Tables B.6-B.10).

Though the results are not statistically significant, the figures show a general pattern that is of interest. Specifically, the treatments seemed to have more positive effects for middle-income and high-income borrowers than for low-income borrowers. For example,

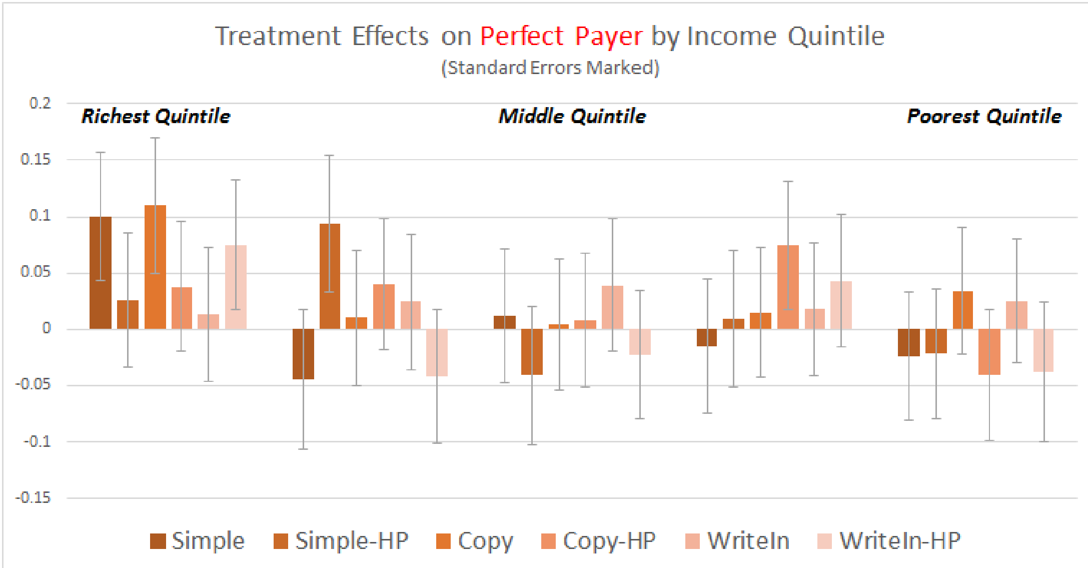


Figure 2.4: *Average Treatment Effects on being a Perfect Payer, by Income Quintile*

Figure 2.3 shows that all treatments except the Simple pledge treatment were associated with small decreases in the likelihood that a poorest-quintile borrower repaid his or her loan, while Figure 2.4 shows that all treatments made richest-quintile borrowers more likely to be perfect payers (with the effects for the Simple and Copy treatments statistically significant at the 90% level; see Appendix Table B.6, specification (2)).

Analysis of the other four outcomes suggest small differential impacts of the treatments by income, in general consistent with the figures—lower income borrowers seem to respond less positively to the treatments than wealthier borrowers. For example, the honor pledges universally decreased the number of days it took richest-quintile borrowers to repay (significant at a 95% level for the WriteIn-HP treatment; see Appendix Table B.6, specification (3)). One might argue that this is because higher income borrowers find it easier to adhere to promises they have made than do lower income borrowers. However, while the patterns in the data are interesting, most effect sizes are not large enough to be statistically significant, meaning that we must be cautious interpreting these results as anything more than suggestive evidence of potential differential impacts of the honor pledges by income. However, these results are certainly not suggestive of any differentially large impact of this intervention on the behavior of the poorest individuals in the subject pool, a pattern we may have expected to see if heightened behavioral biases stemming from poverty were driving a failure to repay loans for the poorest borrowers.

Appendix Table B.11 reports the results from interacting the continuous age variable with the treatments. The results are mixed, with some evidence that the honor pledge treatments were less efficacious for older borrowers. In particular, the results suggest that the treatments (particularly the Simple and WriteIn treatments) were less effective at encouraging older borrowers to be perfect payers (with both coefficients significant at a 90% level; see Appendix Table B.11, specification (2)). However, given the effect sizes and standard errors, there is not sufficient evidence to suggest a strong differential impact of honor pledge treatments by age.

Finally, Appendix Table B.12 reports the results from interacting gender with the treatments. Again, the results are not suggestive of significantly different treatment effects by gender, with mixed results across outcome variables. One potentially interesting result is that the WriteIn and WriteIn-HP treatments were more effective at encouraging loan repayment for men than for women (see Appendix Table B.12, specification (1)), though the effect sizes are not significant. Further analysis is needed on the exact nature of the subject-designed honor codes to better understand whether there is actually an underlying difference, particularly given the insignificance of the interaction coefficients.

2.6 Discussion

Literature in experimental economics and psychology often finds impacts of promises on human behavior and decision making. However, the results of our field experiment suggest small effects from an explicit promise (and indeed, a rather salient promise) on loan repayment behavior in a real-world setting, with money at stake. There are multiple reasons why we might not have observed strong effects in our intervention; we consider these reasons here.

There are three aspects of the experimental design that might explain the lack of an observable effect in the study. First, it could simply be that the experimental interventions were not powerful enough to change behavior. This is a plausible suggestion with this intervention, as with many other behavioral experiments. However, we are relatively confident that our design made the explicit promise quite salient to the borrower—indeed, in the Copy and WriteIn conditions, borrowers were forced to physically type in and sign promise statements, a highly salient and non-trivial process that ought to have heightened any effect of a promise on behavior. So we do not find this argument to be especially compelling as a reason for the absence of significant effects.

Second, it might be argued that the sample size for this intervention was insufficient to detect statistically significant effects. This again is a limitation of this experiment, particularly given that some suggestive patterns appear in the data that are consistent with honor pledge effects on behavior (though the standard errors prevent us from asserting statistical significance). Future work could address this by leveraging larger sample sizes through projects with larger lenders (the firm in this experiment was a startup, which has since grown significantly). Importantly, this experiment provides some ballpark estimates of effect sizes from potential promise interventions in this context, which may inform power calculations for future studies.

Third, it is important to note that the default action for a borrower was to repay the loan—the partner firm's business model involves the collection of bank account information at the moment of loan initiation, to facilitate both loan disbursement and loan repayment when the due date arrives. Therefore, the firm automatically initiates loan repayment on the due date from the borrower's bank account, absent communication from the borrower. This complicates analysis, as it is possible that there may have been subjects who did not wish to repay their loan (despite having sufficient funds to do so), but forgot to contact the firm to cancel the repayment. If this is the case, this might mask underlying behavioral responses to the treatments, by lumping together those who did not wish to repay, but forgot to cancel repayment, with those who wished to repay and actively chose not to cancel repayment.

From an experimental design perspective, there are a number of ways that future work could build and improve upon this study. We highlight three such design tweaks here. First, as mentioned, it would be ideal to replicate this experiment with a larger sample size, to determine if some of the suggestive evidence here is indicative of actual behavioral responses. Second, future interventions could track additional outcome measures, including email opens, time spent on the various treatment pages, and other subject actions to assess whether the interventions had an impact on the cognitive attention of subjects. Third, it would be useful to run such an experiment in a context where individuals needed to make an "active decision" to either repay or not repay, rather than being defaulted into a particular decision.

This experiment and its results also raise several interesting conceptual questions and issues. Some of these relate to the lack of detectable results, while others suggest open questions and promising avenues for future research. We discuss six of these issues briefly here, highlighting their links to alternative hypotheses and potential follow-up research.

First, the simplest explanation for our results is that existing lab experimental work on promise statements in economics do not generalize to real-world decision making.[21] Therefore, in much the same way that List (2003) used an experiment to show that the endowment effect may dissipate in the real world as individuals gain more market experience, this experiment might shed light on the limitations of promises as tools of behavior change. While we find this argument interesting, the lack of other field experimental evidence on promises and their impact in the literature means that we are unable to dismiss their efficacy in the real world.[22] Furthermore, given that we do observe some consistent but statistically-insignificant effects of our treatments relative to the control, we are less certain that our results are evidence that promises "do not work" as a way of influencing behavior. Our experiment only casts doubt on the efficacy of promise statements in the loan

[21]This argument is common in the debate around the validity of behavioral studies in the lab, and is often linked to the fact that subjects in lab experiments tend to be college students, whose decision-making processes may differ significantly from the general population.

[22]One caveat here is publication bias: it is always possible that the lack of field evidence on promises is due to researchers failing to publicize their null results in this area.

repayment context, and provides some evidence that generalizing results from economics and psychology lab experiments to real-world settings is not straightforward.

Second, in this study a promise was made by a client (the borrower) to a firm (the lender), which may have facilitated a "cheap promise" by the borrower. That is, a borrower might have found a promise made to a faceless firm easier to ignore and renege on than a similar promise to an individual or a peer. There is support for this hypothesis in the literature. For example, Karlan *et al.* (2015) finds that repayment reminders are more effective when they included the loan officer's name, while Bicchieri and Lev-On (2007) discuss the difficulty of encouraging cooperation in the absence of face-to-face communication. Further work could explore the nuances of promises by studying the distinction between peer-to-peer and client-to-firm promises. For example, a useful field experiment would be one that randomly varies the information provided about the other individual in a promise scenario, to determine how revealed information about the "other side" of a promise can influence an individual's willingness to follow through.

Third, the promises in this experiment were explicitly private—made by an individual to a firm, with no third-party involvement or social component. In the real world, many promises have a social component, in that you suffer a social or publicly-observable sanction for failing to adhere to your promise. This may explain why, for example, honor codes are so widely used in academic contexts; in these instances, the social stigma associated with violation of the honor code may be as great a deterrent as the punishment for the violation. If this is an important factor in the efficacy of explicit promises, it may explain why the promises in this experiment were not as effective at changing behavior—this experiment relied more on "internal motivation" than "external motivation."[23] A future field experiment could vary the observability of the promise to others in one's social circle, to assess how the social aspect of a promise influences the likelihood that an individual follows through on it. Such an experiment would help disentangle the internal and external motivations underlying a promise, and determine which is more important in influencing real-world

[23]These terms are discussed in section 2.2.1.

behavior.

Fourth, there is an important distinction between non-binding promises and binding commitments. This experiment relied on the former, and it is possible that the latter is much more powerful when it comes to financial behavior specifically. In this context, a promise can be construed as a non-binding agreement, made in the present, to do something in the future that may not be in your future self's best interest.[24] Therefore, a promise is not inherently credible, since you have an incentive to violate it. A commitment, on the other hand, binds you to a course of action (legally or otherwise), and is therefore a more powerful tool of self-regulation. Indeed, Ashraf *et al.* (2006) find that commitment contracts are popular as savings products, suggesting that people might be keen to "bind their hands" in financial decisions, rather than giving themselves the opportunity to get out of a non-binding promise.

In assessing this experiment, we must consider who the subjects in the experiment are—individuals in urgent need of short-term loans. The question we must ask is: to what extent are such subjects swayed by non-binding promises? It is possible that because of the myriad pressures in their lives, individuals in need of short-term loans may be more used to reneging on financial promises (or commitments more generally) than the average person in the population, or may be less likely to take non-binding promises seriously. There are many promising areas for continued research here. For example, a future field experiment on promises among short-term borrowers could focus on how pre-experiment, real-world experiences with agreements (self-reported by borrowers, or using administrative data on past loan repayment if possible) might influence ensuing, post-promise behavior.[25] Alternatively, a longitudinal lab study could explore how prior experience with binding or non-binding agreements could influence future adherence to non-binding promises, by

[24]For an expanded and eloquent description of these issues and terms, see Schelling (1960), to which we owe much of this argument's structure.

[25]This is not possible to do with this experiment's data, since all borrowers were first-time borrowers with the partner firm.

randomizing the amount of previous exposure to such agreements across subjects.[26]

Fifth, there is a real possibility that time plays an important role in mediating the effect of a promise. One model of behavior would be that people will "say anything" in the moment to get what they want, be it a loan or something else, because they are present-biased. This might encourage people to make a promise that they have no intention (or ability) to keep in the time frame of the agreement. In this experiment, the number of days until loan repayment was selected by the borrower, which enabled people to "select" into the repayment time frame that best suited their preferences, and potential biases, around time and money. Another experiment could just as easily randomly assign the number of days until repayment by restricting borrower choice on this dimension. This could help determine how the time between a promise and the time of follow-through might mediate one's likelihood to adhere to a promise, and if so, whether that relationship exhibits time inconsistency.

Finally, the results from this experiment suggest that loan repayment may simply be a behavior that is less manipulable through behavioral interventions—a characteristic we might term "behaviorally inelastic." For example, in this experiment, a subset of subjects may not have repaid their loans, even if they felt a strong internal motivation to do so because of their "honor pledge," simply because they did not have the financial means to repay. In other words, their financial constraints prevented them from changing their behavior. Notably, past studies on promises focused on contexts where "following through" on a promise was always an option comfortably situated in a person's choice set; with academic honor codes, students can always elect not to cheat, while in dictator games, subjects can always send the promised amount of money. In other words, behavior in these instances was potentially more elastic than with loan repayment. Future field experiments that find ways to randomly and compellingly vary "behavioral elasticity" in real-world contexts would be major additions to existing literature. Such experiments would help us

[26]This could also be done using an artefactual or framed field experiment, to use the terminology of the widely-used taxonomy in Harrison and List (2004).

understand the extent to which observed responses to behavioral interventions in real-world settings are due to underlying biases and quirks in human decision making, or simply to the magnitude of behavioral elasticities in the specific experimental context.

Chapter 3

Does Negative Emotion Increase Risk Aversion?: Evidence from Exam Grades and NFL Games[1]

3.1 Introduction

A growing body of work in psychology and behavioral decision research explores the carry-over effects of emotion on economic decision making. Most of these studies are randomized laboratory experiments in which treated participants are exposed to a manipulation that is designed to elicit one particular emotion. For example, Brooks and Schweitzer (2011) induced anxiety in participants using a three-minute audio clip of the theme from "Psycho," Small *et al.* (2006) induced anger and sadness by asking participants to write essays about the September 11th attacks, and Valdesolo and DeSteno (2006) induced positive affect using a five-minute video segment from "Saturday Night Live." However, studies have also used an alternative field approach, in which a quasi-random, real-world event determines the emotional states of subjects. For example, Cunningham (1979), Schwarz and Clore (1983), and Guéguen (2013) link sunshine and weather to study participants' moods, Butler and

[1]Co-authored with Julia Lee and Matthew Ranson

Mathews (1987), Heilman *et al.* (2010), and Krupić and Corr (2014) link students' exam performance to anxiety and/or negative emotions, Hirt *et al.* (1992) measure the impact of sports teams' wins and losses on fans' emotions and self-esteem, and Pierce *et al.* (2015) assess the impact of election outcomes on the happiness and sadness of partisans. A smaller body of research has used these quasi-random variations in emotion to explore the actual behavior of individuals when they experience emotional shocks, exploring the effect of sports results on domestic violence (Card and Dahl, 2011) or voting behavior (Healy *et al.*, 2010; Miller, 2013).

In this paper, we present the results of two field studies we conducted that use the quasi-experimental approach to assess the impact of emotions on risk preferences. The central element of the methodology in our field studies is having participants complete tasks after a real-world event determines their emotional state. Critically, the event must include an element of randomness that allows the researcher to conduct causal inference. The two field contexts in our studies differ in one critical way, however—one involves real-world emotion linked to personal performance and self-responsibility (a student's emotional reaction to academic performance), while the other involves real-world emotion linked to an external force (a fan's emotional reaction to a football game). In both field studies, we find a strong relationship between the real-world event and negative emotions, measured both as specific emotions and using a composite variable for negative emotion constructed using principal component analysis.

The results on the link between emotion and risk preferences are mixed, however. Using a panel instrumental variable approach, we find strong support for the hypothesis that negative emotions reduce risk-taking behavior in the context with high self-responsibility (academic performance), but no strong relationship between emotions and risk preferences in the context with low self-responsibility (football game outcomes). These results hint at a possible explanation for the empirical inconsistencies in the literature linking emotions and risk preferences. Specifically, we posit that in the real world, the link between negative emotion and risk preferences may rely on the extent to which the emotional response comes from personal disappointment or from an external incident, divorced from self-responsibility. In other words, our results suggest that negative emotions may influence risk preferences only when they stem from something an individual did, or could have done.

Our work makes two important contributions to existing literature. First, we provide real-world evidence on the relationship between emotions and risk attitudes, an area that has mostly been explored in controlled laboratory settings. Indeed, our mixed results offer interesting testable hypotheses around self-responsibility's possible mediating role. Second, we offer methodological and analytical insights and techniques that build on the work of existing researchers in this area. In particular, we use both principal component analysis and a panel instrumental variable approach—two techniques that we feel are essential for field research in this area, and an improvement on existing methods.

The first field study was conducted with 114 undergraduate students in a large science class at a selective university. We asked these students to complete a risk-elicitation task immediately after they learned their grade on a midterm exam. In this task, students chose whether to accept or reject a series of small gambles involving real money. Because performance on the midterm exam includes an element of randomness, we argue that whether a student's grade was better or worse than she expected provides a useful source of emotional variation than can be used to measure the carry-over effects of emotions on choices in the risk-elicitation task.

The second field study was conducted with 248 fans of various professional football teams playing in the National Football League (NFL). These subjects completed a variant of the Balloon Analogue Risk Task (BART) in the immediate aftermath of NFL games played in late 2012 and early 2013, again with real money at stake. The task was designed to provide us with a measure of their risk preferences, in a way that could be assessed quickly and remotely—we used email surveys that could be completed on a computer or mobile device. Again, the key identifying assumption is that the results of an NFL game are quasi-random, in the sense that a good or poor outcome in the game is always possible and can cause a genuine affective shock for an emotionally-invested fan.

Based on the results from these field studies, we draw three main conclusions. First, we show that a field design can be used to observe participants experiencing strong emotional responses. In our field studies, students who did poorly on the exam relative to their own expectations and fans whose teams lost reported substantially higher feelings of sadness and anger, and substantially lower feelings of happiness. We argue that the range and intensity of emotion induced by the midterm exam and game outcomes are likely to be much greater than what is possible to observe in a laboratory setting, where practical and ethical concerns typically preclude researchers from inducing strong emotions in subjects. Of course, this advantage comes at a price: it is more difficult to isolate specific emotions using a field design. We propose a methodological approach to deal with this difficulty, as mentioned—using principal component analysis to create a single variable that captures "negative emotion," based on the set of specific emotions reported by each subject.

Second, we find that utilizing a formal statistical model provides important discipline in interpreting the validity of the quasi-experiment. For example, in our first field study, we find evidence that the emotional shock caused by the midterm exam is not unconditionally independent of participants' characteristics. In particular, participants who did worse than they expected on the exam were also more likely to state that school is very important, and to exhibit high baseline levels of anxiety. The only way to control for these violations of the identification assumption is to adopt a panel instrumental variables approach with repeat data from each participant. Unfortunately, in our review of previous studies using quasi-experimental designs to study emotions, we find few studies that conduct this type of test of baseline characteristics or that use an appropriate instrumental variables approach.

Third, our mixed empirical findings lead us to conclude that further research is needed to better understand the potential mediating role for self-responsibility in explaining the relationship between negative emotions and risk preferences. While we hypothesize based on our results that emotional responses to real-world situations with greater self-responsibility may influence risk in ways that low-self-responsibility situations do not, we caution that our findings only support this as a hypothesis—more evidence from follow-up studies, both in the lab and the field, are needed to speak more authoritatively about a potential relationship.

The remainder of this paper is organized as follows. Section 3.2 reviews existing literature on quasi-experimental approaches to the study of emotions, and presents the main hypothesis for our field studies. Sections 3.3 and 3.4 describe our methodologies and analytical approaches in the field studies, and Section 3.5 presents the results of the studies. Section 3.6 provides a discussion and brief conclusion.

3.2 Background and Hypotheses

3.2.1 Quasi-experimental Approaches to Emotions and Behavioral Research

The links between emotions and decision making are complex (Smith and Ellsworth, 1985; Scherer, 1988; Smith and Lazarus, 1993). Previous work has established that emotions affect cognitive processes such as perceptions of risk, valuation, and attitudes (Lerner and Keltner, 2000; Tiedens and Linton, 2001). At the same time, emotions are influenced both by anticipation of the future consequences of a decision and by incidental factors unrelated to the decision or its consequences (Loewenstein and Lerner, 2003). Furthermore, the source of these incidental emotions can be either dispositional affect ("a tendency to react in a particular affective way to a variety of events across time and situations") or situational affect ("a transient reaction to specific events").[2] Because situational affect can be manipulated using experiments, much recent research has focused on how this source of incidental emotion affects judgement and decision making (Cryder *et al.*, 2008; Ariely and Loewenstein, 2006; Han *et al.*, 2012; Lerner *et al.*, 2004).

As described in the introduction, a new quasi-experimental field approach is emerging as an alternative to laboratory studies of emotion. However, the literature utilizing this approach is sparse, and much of the work focuses on how quasi-random variations in real-world settings can influence emotions, and not necessarily how this change in emotional

[2] Loewenstein and Lerner (2003, p. 632)

response translates into actual behavior. We have identified three main sources of real-world variation explored in the literature, and review some of the existing work here.

First, a few authors have studied the effects of variation in weather to explore mood and behavior. For example, Schwarz and Clore (1983) use day-to-day variations in weather as a way to induce positive and negative moods, in order to test how mood affects subjects' judgements about happiness and life satisfaction. In terms of research on behavior, Cunningham (1979) finds that sunshine and temperature affect the number of study questions that subjects are willing to answer, and that restaurant tipping behavior is positively correlated with sunshine. Similarly, Larrick *et al.* (2011) find that high temperatures are associated with greater hostile behavior by baseball pitchers (measured by pitchers hitting batters with pitches). Another area with significant existing research is the effect of weather on financial returns. For example, Schneider (2014) reports on analyses of weather and stock returns, positing that barometric pressure is the best weather predictor of returns in stock markets in London and Frankfurt (with higher barometric pressure associated with higher returns).

A second line of research uses college exams as a source of variation in college students' emotions. For example, Butler and Mathews (1987) studied the effects of anxiety on subjective probability assessments by recruiting college students with upcoming exams. The authors administered questionnaires to the students with exams, as well as to a control group of students without exams, one month and one day before the exam. They found that on the day before the exam, reported anxiety increased, and students gave higher probability ratings to negative events and lower probability ratings to positive events. Similarly, Heilman *et al.* (2010) asked students to complete the Balloon Analogue Risk Task immediately after they learned their final grade in an undergraduate class. They found that cognitively regulating emotions reduced risk aversion associated with emotions, by increasing one's sense of emotional control.

A third set of studies exploits individuals' emotional responses to sports games. For example, Hirt *et al.* (1992) showed subjects live televised men's basketball games, and then evaluated their mood and their performance on several laboratory tasks. Their results show that subjects' feelings of self-esteem depended on whether their favored team won the game, and that subjects' assessments of their own performance on a laboratory task decreased if their favored team lost. A growing literature has also connected sports outcomes to real-world behavior. For example, Card and Dahl (2011) document a 10% increase in male-on-female family violence in the aftermath of upset losses in professional football matches in the United States. Similar work has explored violence related to college football games and international soccer matches (Rees and Schnepel, 2009; Kirby *et al.*, 2014). Researchers have also explored the effect of sports outcomes on voting behavior, generally finding that team success increases the likelihood of incumbent election (Healy *et al.*, 2010; Miller, 2013). Risk has not been widely studied in this context however, though Berument and Ceylan (2012) use international soccer match outcomes to explore stock returns and the return-volatility relationship, finding some evidence that risk aversion increases after losses and declines after wins.

Overall, this literature suggests that field settings are a fertile place for research on emotions and decision making. Unfortunately, however, many of these studies suffer from a shared statistical problem: they all use somewhat inappropriate techniques to calculate the causal effect of emotions on decision making. Furthermore, as we have described, much of the existing work explores only the link between the quasi-random event and the emotional response of the individual, without using this relationship to explore individual decisions. These methodological problems are an important source of motivation for our field studies.

3.2.2 Hypotheses about Emotion and Risk Aversion

Our field studies explore a simple but important research question: how do emotions unrelated to the context of risk itself affect decisions that involve risk? To formulate hypotheses, we draw on several theories of emotion and decision-making.

Early work has examined the effect of either positive or negative affective states on risk preferences. The mood-maintenance hypothesis (Isen and Patrick, 1983) posits that individuals in a positive mood will be more risk averse than those in a neutral mood, because people experiencing positive affect are motivated to maintain their positive state. In other words, positive affect deters individuals from taking high-stake risks, as they have a greater chance of disrupting their good mood. It follows, then, that negative affect would increase risk-taking (to repair one's aversive mood).

Alternatively, Forgas (1995)'s Affect Infusion Model (AIM) proposes that positive affective states would induce risk-seeking behavior, while negative affective states would induce risk aversion. Forgas argues that a positive affective state may help individuals appraise risk positively, while a negative affect state may make it difficult for individuals to accept the negative consequences associated with a risky decision. Several studies support the predictions made by the AIM (Arkes *et al.*, 1988; Chou *et al.*, 2007; Grable and Roszkowski, 2008). For example, Chou *et al.* (2007) found that individuals who watched a movie to induce happiness were more likely to take risks than those who watched a movie that induced sadness.

The mixed results and opposing theories around the link between affective state and risk preferences suggest that there may be important mediating factors influencing the relationship between these variables. We posit that an important determinant of the relationship could be the nature of the event causing the emotional response. Specifically, we focus on the possibility that the level of self-responsibility associated with the event that induces the emotional response may influence whether there is an impact of affect on risk preferences. Before making this case with our field studies and their results, it is important to note alternative explanations for the mixed results in existing literature. We focus on two in particular, before suggesting our own hypothesis related to self-responsibility.

First, some research suggests that affective state might have a differential impact on risk preferences based on the level of arousal (or alertness) induced by the affective state. Mano (1992, 1994) found that individuals in a negative mood took more risks than those in a neutral state, but that this effect was driven by the arousal associated with the affective state rather than the valence (positive or negative) of the affective state. According to Mano's arousal hypothesis, outlined in Mano (1992), heightened arousal limits attentional capacity, thus creating an information-processing bias independent of the valence of the affective state. Several studies support the hypothesis that emotional states associated with high arousal increase risk-seeking tendencies (Fessler *et al.*, 2004; Leith and Baumeister, 1996). These findings suggest that the precise characteristics of the emotions experienced prior to making a risk decision (and not just their valence) may have significant impacts on decisions, consistent with our proposal of a role for self-responsibility in mediating a possible relationship.

Second, the mixed results in existing research on affective states may be due to differences in how similarly-valenced emotions impact risk preferences. For example, although both fear and anger are negatively-valenced affective states, they seem to have opposing effects on risk perception and behavior, due to an underlying difference in cognitive appraisals. Drawing on the Appraisal Tendency Framework (or "ATF;" Lerner and Keltner, 2000), anger is characterized by a heightened sense of certainty and control, while fear is characterized by a lack of certainty and control. Thus, experimentally-induced and dispositional anger produces more risk-seeking choices and optimistic risk perceptions, while fear leads to more risk aversion and pessimistic risk perceptions (Lerner and Keltner, 2000, 2001; Lerner *et al.*, 2004; Lerner and Tiedens, 2006). Similarly, Raghunathan and Pham (1999) found that anxious individuals preferred low-risk/low-reward options, whereas sad individuals preferred high-risk/high-reward options. Similar to fear, anxiety is characterized by feelings of uncertainty and a lack of control, and thus activates a goal of uncertainty minimization. On the other hand, sadness is characterized by the loss of a source of reward, and thus activates a goal of reward maximization. Therefore, if one's negative affective state includes anger, sadness, fear, and anxiety, the effect of discrete emotions may cancel each other out. While we focus on negative affect more broadly rather than specific emotions, the ATF merits consideration in light of our findings (which we touch on in the discussion section).

The ATF and its predictions hint at a key challenge to studying emotions in the field. While emotion theorists have established a clear set of hypotheses about the effects of specific emotions on risk-taking behavior, what is less clear from this previous work is how these emotions might act together to influence risk preferences in the real world. It is often the case that individuals experience combinations of emotions: anger and sadness over a financial setback, or happiness and anxiety about a new relationship, for example. Indeed, scholars have argued that one can experience conflicting emotions at the same time, and such experiences of mixed emotions are natural and frequent (Cacioppo *et al.*, 1997; Larsen *et al.*, 2001). Because the events on which our studies are based could induce any combination of these emotions, we formulate the general hypothesis that in the aggregate, negative emotions reduce risk-taking behavior. To do this empirically, we use principal component analysis, which we discuss further in section 3.4.2.

Importantly, the two real-world contexts that we study involve emotional shocks that differ in an important way—namely, one involves emotion brought on by a personal success or failure, relative to expectations (receipt of a midterm score), while the second involves emotion in response to an external event over which the subject had no control (a football game). In particular, receiving poor test scores may be attributed to one's own sense of self-responsibility or competence, potentially posing a greater threat to one's self-esteem than an NFL game outcome. Existing research suggests that individuals with lower self-esteem are more likely to prioritize self-protection by minimizing the negative consequences of their actions, while those with higher self-esteem are more likely to make risky decisions (Brockner *et al.*, 1983). In line with this argument, higher self-esteem has also been linked to riskier monetary decisions (Josephs *et al.*, 1992). We extend this logic to our field studies, and suggest that the effect of naturally-occurring negative emotions on risk preferences may be more pronounced in real-world scenarios where the event causing the emotion is more easily linked to an individual's own actions.

As noted in the introduction, much of the past research on emotions and risk induced either mood or discrete emotions using video clips, which cannot be linked to self-responsibility. Thus, these studies do not test the extent to which personal control over the event causing the emotional response plays a role in influencing subsequent risk decisions. While our studies do not randomly vary the perceptions of self-responsibility from the events, relying instead on the natural variance between the contexts, they do allow us to indirectly test the role of attributions of responsibility in mediating the relationship between negative emotions and risk preferences.

3.3 Methodology for Field Studies

3.3.1 Study 1: Evidence from Midterm Exam Performance

Our first field study uses the emotional shock from the receipt of a midterm exam grade to assess the impact of real-world emotion on risk attitudes. This field approach has been used in the past (Butler and Mathews, 1987; Heilman *et al.*, 2010; Krupić and Corr, 2014), but we build on existing work by refining the method for causal estimation in two ways. First, we incorporate principal component analysis to reduce dimensionality in the dependent variable, negative emotion, and address the problem of multiple simultaneous emotions. Note that the latter issue is especially prominent in field settings, which lack the level of control over specific emotions that laboratory studies can achieve. Second, we use an instrumental variable approach to refine the independent variable, isolating the variation in emotion caused by exam performance specifically.

Participants & Recruitment

Subjects for this study were undergraduate students recruited from a large science class at a highly-selective university. We chose this science class because it is an important requirement for students who are interested in obtaining a post-graduate degree. A few weeks into the semester, the course professor (acting on our behalf) informed students before a lecture that they could voluntarily sign up for a "decision-making study" in exchange for cash payments and the possibility of winning a $100 gift card to an online retailer. Students were told that participation would entail staying after class on two separate days in the future to complete a short survey. At this point, a total of 114 students signed up for our study.

Of the 114 recruited students, 94 students actually stayed after class and participated in our first survey, and all but one of these students completed the second, follow-up survey online. All participants were paid $20 guaranteed for the completion of the two post-class surveys, and made additional money based on their decisions in the tasks in the two surveys they completed. The range of additional payments for the decisions made in the two surveys was $2 to $44.50. Additionally, one subject who completed both surveys was selected at random to receive a $100 gift card to an online retailer.

Study Procedure

The first of the two surveys in the study, the "treatment" survey, was administered on the day that students learned their grades in the course's midterm exam. At the end of lecture, the teaching staff returned students' graded exams. Immediately afterward, students who had previously signed up for the study completed a paper survey that began with three decision-making tasks.[3] One of the tasks was a risk-elicitation task, presenting three gambles that involved choosing between $2 of real money or a risky prospect (e.g., a 25% chance of winning $6). After completing these tasks (but prior to payment), participants completed a short questionnaire that asked for: 1) basic demographic characteristics; 2) current and usual emotional state (specifically, happiness, anger, anxiety, and sadness); 3) their performance on the midterm exam (0-100); and 4) some questions about the midterm and their performance, including a self-assessment of their performance relative to their expectations (on a 1-9 scale, from "much worse" to "much better"). Participants were paid online within 2-3 days of completing the treatment survey, based on their decisions in the survey tasks (one of the three risk-elicitation tasks was chosen at random for payment, along with a dictator game task).

[3]See online appendix for the treatment and follow-up surveys. Three decision tasks were used in each survey, namely: 1) a dictator game that involved dividing $7 of real money between oneself and an anonymous partner; 2) a "filler" task that asked students to rate several consumer items; and 3) a risk-elicitation task. The order in which these three parts were presented was randomly varied across participants, though the filler task was always the second of the three tasks. However, the analysis of all tasks is beyond the scope of this paper—we only use the data from the risk-elicitation task.

Table 3.1: *Test Score Study: Risk-Elicitation Tasks (Treatment and Follow-Up Surveys)*

	Treatment Survey	**Follow-Up Survey**
Risk Task #1	\$2 for sure, or 10% chance of \$20	\$2 for sure, or 10% chance of \$25
Risk Task #2	\$2 for sure, or 25% chance of \$6	\$2 for sure, or 25% chance of \$7
Risk Task #3	\$2 for sure, or 20% chance of \$15	\$2 for sure, or 20% chance of \$14

The second survey in the study, the "follow-up" survey, took place roughly four week later. All participants who had completed the treatment survey were emailed a link to an online questionnaire that they could complete at their convenience. This follow-up survey contained questions and tasks that were similar to those in the treatment survey. However, to reduce potential response bias from cognitive dissonance reduction across the two surveys, the follow-up survey included several minor differences. For the risk elicitation task, the probabilities and payoffs for the uncertain gambles were modified slightly, but were generally similar to those in the treatment survey. Table 3.1 outlines the specific probabilities and payoffs used in the risk elicitation tasks in both the treatment and follow-up surveys.[4]

3.3.2 Study 2: Evidence from NFL Fans

Participants & Recruitment

A total of 248 NFL fans were recruited from online sports message boards, using approved postings soliciting participation.[5] These solicitations included a link to an online pre-screening survey that enabled us to determine the eligibility of potential participants, based on: 1) whether they completed the entire pre-screening survey upon recruitment; 2) how committed they were to their football team; 3) how available they were to complete post-game surveys; and 4) whether they had a PayPal account to receive payments. Based on this pre-screening, we sent out email invitations to 203 fans, asking them to participate

[4]The changes for the other tasks were as follows: 1) in the dictator game, subjects divided \$6.50 instead of \$7; and 2) the "filler" task asked about a different set of consumer items.

[5]Two message board websites were used for recruiting: 1) http://www.thefantasyfootballguys.com; and 2) http://forums.footballguys.com

in a baseline survey for payment at their convenience, followed by further surveys in the aftermath of upcoming NFL games. Participants had to complete the baseline survey in order to be part of the rest of the study.[6] Of the 203 fans invited to complete the baseline survey, 163 completed the survey and remained in the study going into the first week of NFL games.

All participants were paid a basic $5 show-up fee via PayPal after completing each survey, including the baseline survey. In addition, they had the opportunity to win up to an additional $26 in each survey, based on their choices in the risk elicitation tasks (the dollar values varied slightly depending on the payoffs and probabilities that we used for each survey task).

Study Procedure

The 203 selected participants were sent an email with a link to a baseline survey, which included more information about the study, solicited contact and basic demographic information from participants, and required participants to complete a risk-elicitation task modelled on the Lejuez *et al.* (2002) Balloon Analogue Risk Task (BART). In order to measure the baseline level of risk-taking and to familiarize participants with the rules of the BART, all participants played three distinct "games" of the BART at baseline, for real money.

The task worked as follows: each distinct BART game began with participants being presented with a virtual balloon, which had an initial dollar value to participants (either $1 or $2, as shown in Table 3.2). During each of up to seven rounds of the game, the participant chose to either "inflate" or "not inflate" the balloon. If the participant inflated the balloon successfully, he or she could increase the dollar value of the balloon (by either $1 or $2, as shown in Table 3.2) . However, there was a small probability (known to the participant) that adding air to the balloon would make it pop. If the balloon did pop, the participant did not earn any money from that game. If the participant chose to stop inflating the balloon at any

[6]How many post-game surveys a given participant completed depended on how many games his/her favorite team played in the postseason.

Table 3.2: *NFL Fans Study: BART Details by Survey Week*

	Baseline	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Initial Balloon Value	\$1	\$1	\$2	\$2	\$2	\$2	\$2
Probability of Popping | Inflate (Game 1)	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Probability of Popping | Inflate (Games 2&3)	0.25	0.25	0.5	0.5	0.5	0.5	0.5
Increase in Value | Inflate Success (Game 1)	+\$1	+\$1	+\$1	+\$1	+\$1	+\$1	+\$1
Increase in Value | Inflate Success (Games 2&3)	+\$1	+\$1	+\$2	+\$2	+\$2	+\$2	+\$2

point before the balloon popped, the game ended and the participant got to keep the money he or she had earned up to that point.

Each participant was then invited to complete up to six similar "post-game" surveys, after NFL games in late 2012 and early 2013.[7] The day before each game, we sent participants an email message reminding them that they would recieve a short survey to complete immediately after the football game. Then, at the exact moment that their favorite team's game ended, we sent each participant a text message and an email with a link to the post-game survey. The survey consisted of a few short questions about participants' emotions and how they felt about the football game they just watched, along with three BART games, played for real money. Participants were given a 20 minute window after the game ended to fill out the post-game survey. This ensured that their emotional response to the game was still strong at the time of survey completion.

The payoffs and probability of popping varied slightly across BART games across weeks. This was to ensure that we had sufficient variation in inflation decisions—we noticed after Week 1 that participants were inflating balloons at higher rates than anticipated. To adjust our design to encourage less inflation by participants, from Week 2 onward, we altered

[7]This period covered the NFL playoffs and Super Bowl, when multiple teams were no longer playing, so there were multiple weeks where the possible number of participants was much less than the initial 163 participants.

the initial value of the balloon in each game (increasing it from \$1 to \$2), and changed the payoff structure for two of the three BART games each week (games 2 and 3). Specifically, for these games we raised the probability a balloon would pop if inflated from 25% to 50%, while simultaneously increasing the payoff for inflating a balloon successfully in any given decision from \$1 to \$2. Note that game 1 in each week had the same structure throughout the study (inflation had a 25% pop probability, and an incremental \$1 payoff). Table 3.2 outlines the differences in the BART game setup across survey weeks.

3.4 Analytical Approach

Unlike a laboratory experiment, in which participants are explicitly randomized into treatment and control groups, our field studies rely on naturally-occurring emotional variation. As a result, the statistical validity of the results depends heavily on whether it is possible to identify a source of emotional variation that is unrelated to participants' characteristics. We focus on the identifying assumptions that conditional on an individual-specific fixed effect: 1) in Study 1, a participant's absolute midterm performance and midterm performance relative to expectations are orthogonal to any other variables that changed between the treatment and follow-up surveys; and 2) in Study 2, game outcomes and team performance are unrelated to participant characteristics.

3.4.1 A Simple Model

To formalize the intuition for these assumptions, consider the following simple model of how risk preferences depend on emotional state, using Study 1 as an example for illustrative purposes. Let R_{it} represent individual i's risk-taking behavior at time t. Suppose that R_{it} is a function of current emotional state E_{it}, an unobserved individual-specific personality characteristic γ_i, an unobserved factor Z_{it}, and an error term ϵ_{it}:

$$R_{it} = \beta_1 E_{it} + \beta_2 Z_{it} + \gamma_i + \epsilon_{it} \tag{3.1}$$

Suppose also that emotional state itself depends on midterm score relative to expectations M_{it}, an individual-specific personality characteristic γ_i, the unobserved factor Z_{it}, and an error term ξ_{it}:

$$E_{it} = \alpha_1 M_{it} + \alpha_2 Z_{it} + \gamma_i + \xi_{it} \tag{3.2}$$

Now, the central challenge in trying to estimate β_1, which represents the effect of emotions E_{it} on risk-taking behavior R_{it}, is that E_{it} is correlated with both omitted variables, γ_i and Z_{it}. For example, in Study 1 it seems possible that both emotion and risk preferences could be correlated with personality characteristics such as stability or intelligence. Furthermore, it is possible that some time-variant omitted variable (e.g., the time of day at which each participant chose to complete the online follow-up survey) may influence both emotion and risk preferences. Thus, due to the omission of unobserved correlated variables, using cross-sectional or panel data to estimate Equation (3.1) may result in a biased estimate of β_1.

However, if—conditional on an individual-specific fixed effect—performance relative to expectations (M_{it}) is orthogonal to the unobserved variable Z_{it}, then β_1 can be estimated using a panel instrumental variables approach that includes individual fixed effects in both stages. We use this statistical approach in the results section, for both field studies. For comparison, we also present results from "naive" cross-sectional regressions of risk preferences on emotions.

3.4.2 Outcomes and Methodological Approaches

There are two important components of our analytical approach in this paper, which build on and extend existing work on emotions. First, for both field studies, we adopt an empirical approach that factors in the complexity of real world emotions. One of the challenges of studying emotions in the field is that it is difficult to isolate the effect of individual emotions on risk preferences. Thus, in the analysis, we develop a composite "negative emotion" variable based on participants' self-reported levels of current emotional state. To construct this variable, we first use principal component analysis to identify the main component of common variation between the specific emotion variables collected in each field study.

Appendix Tables C.2 and C.10 present the results of the principal component analysis for the two field studies. We interpret the component with the largest eigenvalue as a natural interpretation of negative emotion, including roughly equal amounts of negative emotion in the data. We then project each participant's emotion onto this principal eigenvector. This process reduces the dimensionality of the emotions data while still preserving the maximum possible amount of emotional variation.

In addition, we propose and attempt a unique approach to estimating risk that goes beyond those used in existing work using the BART to study risk. Specifically, we embed our data into a simple structural model of risk preferences in order to estimate a lower bound for individual risk aversion, which we then use as a dependent variable in the analysis.

Study 1: Midterm Exam Performance

In Study 1, participants make three distinct decisions over risk gambles in each of the two surveys (treatment and follow-up). In each instance, the participant chooses between \$2 for sure, or a known gamble. We therefore compute a "riskiness" outcome variable as the percentage of times in each survey that a participant chose a risky decision, preferring the gamble to the guarantee of \$2. While this is not a precise estimate of risk preferences, we argue that it serves as a reasonable proxy, particularly since we have repeat observations for each individual in the data.

We analyze the relationship between risk and emotion using our data in two ways. First, we conduct a simple "naive" regression of riskiness on the negative emotion variable we construct using principal component analysis. We do this both with and without panel data, demographic controls, and individual fixed effects. Second, we conduct a two-stage instrumental variables regression, using both midterm score (0-100 scale) and a self-assessment of midterm score relative to expectations (1-9 scale) as instruments for negative emotion. Importantly, we adopt a panel instrumental variable approach, using time and individual fixed effects in both stages of the instrumental variable regression.

Study 2: NFL Fans

In Study 2, participants make a variety of ever-changing decisions about risk and reward across BART games—as the balloon inflates, the potential loss from popping increases, while the return to successfully inflating the balloon stays constant. Existing studies using the BART have used a number of different proxy measures for risk preferences based on decisions in the BART, including the number of unpopped balloons, total pumps, adjusted pumps (the average number of pumps on all balloons that did not pop), and the maximum number of pumps on a balloon in a given set of games. As Pleskac *et al.* (2008) and DeMartini *et al.* (2014) have noted, many of these proxy measures are biased because they do not adequately factor in popped balloons—when a balloon pops we cannot observe the counterfactual of what a participant "would have" done if the balloon had not popped. This is an inevitable issue with the BART. Nevertheless, we show results using two of the more common outcome variables from the BART from existing literature, adjusted pumps and total pumps in each week.

However, we also use a third outcome variable, which we believe represents a potential improvement on existing practice using the BART task. Specifically, we construct a "risk aversion" variable using a structural model of risk preference, and use it as a dependent variable in the analysis. This risk aversion variable is defined as follows:

$$riskaversion = min(\frac{((1-\alpha)W_w + \alpha W_L) - W_Q)}{W_L^{0.5}}) \tag{3.3}$$

In this expression, α is the probability that the balloon pops in a given decision, W_W is the participant's total wealth if she successfully inflates, W_L is the participant's total wealth if she unsuccessfully attempts to inflate, and W_Q is her wealth if she quits without attempting to inflate. We use the minimum value observed for each participant in a given week (covering three BART games), in order to identify the value at which the respondent switches from the risky option (inflating the balloon) to the safe option (deciding not to inflate). At this point, the respondent's certainty equivalent from the gamble is equal to W_Q. We then construct a measure of risk aversion by calculating the difference between the expected value of the gamble and the respondent's certainty equivalent. Positive values of this variable indicate risk aversion; zero indicates risk neutrality; and negative values indicate risk seeking. To convert this difference into a measure of the local curvature of the participant's utility function, we normalize by dividing by the square root of the participant's wealth in the losing state of the world. This normalization captures the intuitive pattern that people become more risk averse when gambles are scaled up by a constant.

By conducting the analysis with both our constructed risk aversion variable and two existing, commonly-used BART risk variables in the literature, we can assess the robustness of our new approach and offer methodological insights for future researchers using the BART game.

We analyze the relationship between risk and emotion using our data in a similar way to Study 1. First, we conduct simple "naive" regressions of the risk outcome variables on the negative emotion variable, with and without individual and survey week fixed effects. Note that the specific emotions measured in Study 2 overlap with those in Study 1, but do not match exactly due to the differing characteristics of each real-world setting (Study 2 uses anger, disappointment, sadness, nervousness, happiness, and excitement). Second, we conduct a two-stage instrumental variables regression, using team victory or loss as an instrument for negative emotion. We again use a panel instrumental variable approach, using individual and survey week fixed effects in both stages of the instrumental variable regression.

3.5 Results

3.5.1 Study 1: Evidence from the Classroom

Summary Statistics

A total of 93 students participated in both surveys. However, three of these students skipped important questions, and so in the remainder of the analysis we focus on the sample of 90 students who provided complete responses. Appendix Table C.1 summarizes these participants' characteristics. The table shows that 67.8% of participants were female and that their average college GPA was 3.48. The average midterm score was 74.7 points out of 100 possible, although the distribution has a long left tail.

Appendix Table C.1 also summarizes participants' choices in the treatment and follow-up surveys. Specifically, the bottom two rows of the table describe the main outcome variable for our analysis: the percentage of times the risky option was selected in the risk elicitation task. The table shows that participants chose the risky option 43.9 percent of the time in the treatment survey, and 41.5 percent of the time in the follow-up survey.

Appendix Table C.2 presents the results of the principal components analysis for the development of a single "negative emotion" variable, using self-reported anger, sadness, anxiety, and negative happiness. To check whether this process captures the main empirical patterns in the data, the panels in Appendix C.1 plot the correlations between each of the four basic emotions that we study. The table confirms the results of the principal component analysis: there is a strong, linear, 1:1 relationship between most pairs of emotions. The panels in Appendix C.2 plot the distribution of the negative emotion variable, both overall and broken down by survey version (treatment or follow-up). The negative emotion variable has a mean of 2.96, a standard deviation of 3.19, and a range of -2.75 to 11.62 in the data.

Evaluation of Midterm Performance as an Instrumental Variable for Emotion

We begin our analysis by assessing whether midterm exam performance and self-assessed performance relative to expectations are likely to be appropriate instrumental variables for negative emotions. There are two general requirements that must be satisfied. First, exam performance and performance relative to expectations must have strong, statistically significant effects on negative emotion. Second, exam performance and performance relative to expectations must not be related to any unobservable respondent characteristics that also influence risk preferences (in other words, exam performance must affect the dependent variable only through negative emotion, not directly).

Appendix Table C.5 presents evidence about the first requirement. The table shows the results of OLS regressions of negative emotion on midterm performance and performance relative to expectations. Specifications (1) and (2) show cross-sectional results based only on the responses to the treatment survey. These columns show a consistent pattern: both raw midterm score and performance relative to expectation have strong, statistically significant effects on negative emotion. Specifications (3)-(5), which present panel regressions based on data from both the treatment and follow-up surveys, show similar patterns. What is striking about the results from specifications (4) and (5), however, is that these estimates include dummy variables for each individual participant. Thus, these regressions show that even after controlling for all fixed individual-level personality characteristics, exam performance relative to expectations has a strong, highly significant effect on negative emotion.

Because it is possible that the midterm results might have larger impacts on some emotions than on others, Appendix Table C.4 presents the results of similar regressions of each specific emotion (happiness, anxiety, anger, and sadness) on midterm performance. This table largely confirms the results from Appendix Table C.5. Doing poorly on the exam, relative to expectations, causes students to report significantly higher feelings of anxiety, anger, and sadness, and significantly lower feelings of happiness.

Unfortunately, because many participant characteristics are unobservable, it is not possible to conduct a definitive test of the second IV requirement that exam performance only influences risk preferences through negative emotion. However, it is possible to test whether relative exam performance is related to observable characteristics. Appendix Table C.3 presents relevant evidence. The table shows the results from nine distinct regressions, each of which regresses a participant characteristic on midterm score and midterm performance relative to expectations. The results raise some concerns. There is strong evidence that participants' absolute midterm scores are related to their college GPAs. Furthermore, subjects' performance relative to their own expectations is weakly (and inversely) related to their beliefs about the importance of the midterm and positively correlated with their self-reported levels of baseline anxiety. Similarly, there is a weak, positive correlation between midterm score and perception of midterm importance. These results suggest that a simple cross-sectional instrumental variables approach may be inappropriate, as exam performance may be correlated with unobservable variables that also plausibly influence risk preferences—like ability. Thus, our preferred specifications use a panel instrumental variable approach, which includes individual and time fixed effects.

Emotions and Risk Preferences

The results from the previous section suggest that midterm performance relative to expectations has a strong impact on negative emotion, and that conditional on individual and time fixed effects, an instrumental variables approach may be appropriate. Thus, using exam performance as an instrument, this section presents instrumental variables estimates of the impact of negative emotions on risk preferences.

Appendix Table C.6 presents the main results of this analysis, using a variety of specifications. Specifications (1)-(4) show the results of "naive" regressions of the fraction of risky choices selected on the "negative emotion" variable, with specifications (2)-(4) presenting results using data from both the treatment and follow-up surveys with various controls. The coefficient on negative emotion in specification (4), using individual fixed effects, suggests that a one point increase in negative emotion causes a participant to choose, on average, 1.12 percentage point fewer risky gambles. The effect is not significant, however.

Specifications (5)-(9) present estimates based on an instrumental variable approach, with specifications (8) and (9) representing our preferred specifications, using a panel instrumental variables approach with individual and time fixed effects. Specification (8) uses performance relative to expectations as an instrument for negative emotion, while specification (9) uses both midterm score and performance relative to expectation as instruments. These columns show a significant negative relationship between negative emotion and risk-taking behavior, with coefficients that imply that a one point increase in negative emotion causes a 3.8 to 4.3 percentage point decrease in the number of risky options selected.

To illustrate the instrumental variables results, the visuals in Appendix C.3 plot emotions and risky choices against performance relative to expectations. To construct the two panels in Appendix C.3, we first-differenced each of the two dependent variables (negative emotion and percent risky choices) and then regressed these differences on first-differenced absolute midterm score and a first-differenced time period variable.[8] We then plotted the residuals from these regressions against midterm performance relative to expectations. Panel (C.3.1) shows the result for residual negative emotions, and panel (C.3.2) shows the results for residual percentage of risky choices. The results in panel (C.3.1) show that performance relative to expectations has a strong and consistent inverse relationship with negative emotion. In other words, there is evidence that participants who did better than they expected also felt better as a result. Meanwhile, panel (C.3.2) shows a generally positive relationship between performance relative to expectations and the percentage of risky choices—the better participants performed relative to expectation, the more risks they took. Together, these two panels support the conclusion that negative emotion decreases risk-taking behavior.

3.5.2 Study 2: Evidence from NFL Fans

Summary Statistics

While 248 fans were recruited for the study, only 163 of these fans completed the baseline survey for entry into the final pool for eligibility for post-game surveys. Of these 163 baseline participants, 154 completed at least one post-game survey, while the remaining 9 participants stopped responding to survey solicitations after the baseline survey. Panel (C.4.1) in Appendix C.4 provides a visual of the distribution of post-game surveys completed by participants in the study, while Panel (C.4.2) shows a histogram of the number of participants per week. Note that not all 154 active participants completed the post-game survey in the first game week; five participants were added to the study in the middle of the game weeks, because of a delayed baseline response, and only began completing post-game surveys after the third study week.

[8] Note that first-differences and fixed effects regressions produce identical results when there are only two time periods.

Appendix Table C.7 summarizes the demographic and personal characteristics of the fan participants, both in total and disaggregating by team quality. For the latter designation we separate teams into three categories based on their playoff chances entering the first week of the study, the penultimate game week in the NFL season. These categories are: 1) "Playoff" teams, who were assured a playoff position entering the study (55 fans); 2) "Chance" teams, who had a chance to make the playoffs entering the study (71 fans); and 3) "Out" teams, who had no hope of making it to the playoffs entering the study (37 fans). The table shows that only 3.1% of participants were female, and most were older than 30 and relatively well educated (with nearly 80% having completed college or a post-graduate degree). Also, most fans reported regularly feeling emotional shocks when their team lost, with disappointment and frustration being the most prominent personal emotions felt after a loss.

Appendix Tables C.8 and C.9 show the results when various observable fan characteristics are regressed on dummy variables for the team quality variables described above, to determine if there were statistically significant differences between fans by team quality in our sample. Our results show that fans of the more successful teams in our sample were less likely to be college graduates, and less likely to watch their teams play live than fans of the less successful teams. This might be an indication of "fair weather fan" bias among successful team fans in the sample, where fans of lower quality teams are more likely to be "true" fans (committed to watching their team so late in the season despite their poor performance). However, this seems more likely to be a statistical anomaly, as we see no significant differences in other relevant variables, including "fan level" (a self-reported measure of fan commitment) and self-reported emotions after a team's loss. So we are reasonably confident that fan characteristics are not correlated with team quality in a way that complicates analysis.

As with the first field study, we develop a composite "negative emotion" variable based on participants' self-reported emotions (excitement, happiness, sadness, nervousness, anger, and disappointment) in the post-game period, using principal component analysis. Appendix Table C.10 reports the results of this analysis. Additionally, the panels in Appendix

C.5 show the distribution of the negative emotion variable, both overall and broken down by whether or not a survey was completed at baseline, after a team win, or after a team loss. The negative emotion variable has a mean of -3.46, a standard deviation of 66.31, and a range of -87.4 to 136.4 in the data.

Finally, the panels in Appendix C.6 present the distribution of the "risk aversion" variable, constructed using the structural model of risk presented in section 3.4.2. Again, this is presented overall, and broken down by whether or not a survey was completed at baseline, after a team win, or after a team loss. The risk aversion variable has a mean of -0.25, a standard deviation of 0.27, and a range of -1.8 to 0.2 in the data.

Evaluation of Game Outcome as an Instrumental Variable for Emotion

We begin by assessing the extent to which game outcomes serve as a useful instrument for emotion. As in the test score study, two general requirements must be satisfied. First, game outcomes must have strong, statistically significant effects on negative emotion. Second, game outcomes must not be correlated with any unobservable respondent characteristics that also influence risk preferences (in other words, game outcomes must affect risk preferences only through negative emotion, not directly).

To test the first requirement, we run regressions of specific self-reported emotion variables on team victory or defeat, using baseline survey responses as the omitted comparison group. Appendix Table C.11 reports the results. Note that the table includes two columns per emotion variable—one with and one without individual fixed effects. What is immediately clear from the table is that game outcomes are strongly and consistently associated with large emotional responses, with victories resulting in positive emotional responses and losses resulting in negative emotional responses. Appendix C.7 provides additional, visual evidence. Appendix C.7 presents eight panels, plotting an individual's various reported emotions against the score difference in the game they had watched (the score of their favorite team, minus the score of the opposing team). We see very strong visual evidence of a discontinuity in all emotions (except nervousness) at the point where score difference is zero. In other words, these visuals, along with the results in Appendix Table C.11, provide clear evidence that specific emotions are influenced by game outcomes.

Additionally, we run a regression of the constructed variable for "negative emotion" on team victory or defeat, both with and without individual fixed effects, and report the results in Appendix Table C.13. Consistent with the results in Appendix Table C.11, we find very strong relationships between game outcomes and negative emotion—with team wins reducing negative emotion and team losses increasing negative emotion. Panel (C.7.8) in Appendix C.7 also provides visual evidence supporting the same conclusion. Specification (2) in Appendix Table C.13, which uses individual and week fixed effects, is used as the first-stage regression in the two-stage panel instrumental variable analysis, discussed later in this section.

The second requirement is harder to test empirically. We offer one approach in Appendix Tables C.8 and C.9, regressing participant characteristics on team quality prior to the study. Ex-ante, we would hope that team quality was not correlated with fan characteristics—if it was, it might be a sign of potential bias in our IV approach from possible correlation between game outcome and unobservable fan characteristics that influence risk preferences. Our results in Appendix Tables C.8 and C.9 suggest that while participants are similar across team quality along most observables, fans of worse teams are more likely to be college graduates and more likely to watch their favorite teams play live. This is a potential concern, and suggests that a panel instrumental variable with individual fixed effects is again justified.

Despite this concern, we are more confident in the second IV requirement in this case than in the test score study, due to the nature of the event causing the emotional response. In the test score study, there was a more plausible channel through which relevant unobservable participant characteristics could correlate with exam performance, since the latter was arguably a function of unobservable variables that might influence risk preferences (such as student ability). In this study, however, game outcome was outside the fan's locus of control, making it less plausible that it would correlate with risk-relevant fan characteristics.

Nevertheless, we use a panel instrumental variables approach in the analysis, which should help control for possible violations of the second IV requirement.

Game-Induced Emotions and Risk Preferences

The previous section supports the use of game outcome as a valid instrument for negative emotion, particularly if a panel instrumental variable approach is used. This section presents an assessment of the relationship between emotions and risk preferences, using both "naive" and instrumental variable methods.

We present results using three different outcome measures of risk. First, we use two of the most common outcome variables in the literature using the BART: total pumps and "adjusted pumps" (both computed at the participant/week level). The former measure constitutes the total number of times a participant elected to inflate a balloon in a given week, and the latter represents the average number of pumps in all unpopped balloons for a participant in a given week. Note that if a participant popped all of their balloons in a given week, the adjusted pumps variable would be undefined, and thus that week of participant data would be excluded from ensuing analysis. Both of these standard outcome variables have clear limitations, as discussed earlier, so we also report results using the risk aversion outcome variable we constructed, as described in section 3.4.2.

Appendix Table C.12 presents the results of naive regressions of the three risk outcome variables on team victory, without an instrumental variable approach. Our preferred specifications in this table are specifications (2), (4), and (6), which include participant and week fixed effects (the latter being necessary here because of alterations to the BART rules in later weeks of the study). The results for the total pumps and adjusted pumps variables, presented in specifications (2) and (4), suggest that risk taking was lower during post-game weeks than in the baseline survey, but do not suggest any clear effect of team victory on risk outcomes. Similarly, specification (6) finds no statistically significant relationship between game outcomes and the constructed risk aversion variable. The absence of a link between game outcomes and risk preferences in the naive regressions is depicted visually in

Appendix C.8, which shows the relationship between score difference (favorite team score minus opposing team score) and the three risk outcome variables. It is evident in the three panels in Appendix C.8 that there is no noticeable discontinuity in risk outcomes at the point where score difference is zero, suggesting no real relationship between game outcomes and risky decisions.

Appendix Tables C.14-C.16 present the main results of our analysis exploring the link between emotion and risk preferences, for each of the three risk variables. Each table begins with two "naive" specifications, with and without fixed effects. Our preferred specifications in each table are specifications (3) and (4), however, which use a panel IV approach (using the regression from Appendix Table C.13, specification (2), as a first-stage IV regression).

In all three sets of results, both the naive and IV panel regressions of risk outcome on negative emotion do not show any significant relationship when fixed effects for week and individual are included. For example, Appendix Table C.16 presents the results using the risk aversion outcome variable, constructed using the structural model. Specification (2) in Appendix Table C.16 shows that a one unit increase in negative emotion is associated with a 0.000182 decrease in risk aversion in a naive regression, which is statistically indistinguishable from zero. When a panel IV approach is used, the effect increases marginally, to a 0.000233 decrease in risk aversion, but this is again not significant. Recalling that the risk aversion variable had a mean value of -0.25 and a standard deviation of 0.27 in the data, and that the negative emotion variable ranged from -87.4 to 136.4 in the data, it is clear that the magnitude of the observed effects are not practically significant either.

We conclude based on these results that while NFL games induced very strong emotional responses in participants, there was no strong link between those emotions and risk decisions. The fact that this held true for all three outcome measures for risk strengthens our confidence in the results. Our results also make clear the importance of proper estimation methods—naive regressions (which were effectively comparisons of means) often showed relationships between emotions and the risk decisions of fans, but the effects disappeared when appropriate econometric methods were used.[9]

3.6 Discussion and Conclusions

Our two field studies come to different conclusions about the relationship between negative emotion and risk preferences. On the one hand, our field study on undergraduate students receiving exam scores finds relatively strong support for the conclusion that negative emotions increase risk-aversion. However, our field study with NFL fans paints a different picture, finding no link between negative emotion and risk preferences. What could account for the difference in these results?

One possibility is that the results differ because the two studies are not methodologically identical. The test score study was conducted partly in person and partly online, while the NFL fans study was conducted entirely online and through text messages. The two studies also used different methods for eliciting risk preferences, with the test score study using risk choices over gambles and the NFL fans study using the BART. The subjects for the studies were also very different—the test score study participants were undergraduates, while nearly 80% of the NFL fans studied had at least a college degree. It is possible that these differences could have contributed to the differing results. However, it is worth noting that some of these differences are inevitable given the methodological approach of using quasi-random, real-world events to study emotion—adopting this strategy requires sacrificing some level of control relative to what is possible in the laboratory.

That said, we propose that the difference in results may be largely driven by the nature of the events causing the emotional response. Specifically, while test scores directly reflect on an individual's effort and achievement, the performance of one's favorite sports team does not. This difference in "self-responsibility" for the emotion-inducing event might mean that in the test score study, negative emotion had a greater impact on risk because it was associated with participant self-esteem. This in turn could have increased participants'

[9]See specifications (1) and (3) in Appendix Tables C.14-C.16 for these naive regressions.

desire to self-protect by avoiding risk, in line with some existing research (Brockner *et al.*, 1983). This conclusion would also be consistent with the AIM model (Forgas, 1995), but with an added role for self-responsibility in mediating the link between negative emotion and risk.

We do not believe, however, that sports outcomes are unrelated to self-esteem. Indeed, Hirt *et al.* (1992) finds a relationship between game outcomes and fan self-esteem, and we find very large emotional responses to game outcomes in our data. Instead, it may be an issue of the relative scale of self-esteem's role. Not all emotional responses are the same in the real world, in ways that influence how emotions and decision making interact. The negative emotion we feel when we fail an exam, or lose our job, is quite different than the negative emotion we feel when our favorite TV show is cancelled, or when our beloved baseball team loses the 7th game of the World Series. Therefore, while the magnitude of feeling might be comparable, the underlying connection to our sense of self may be very different. We hypothesize, based on our results, that when emotions come from events outside the realm of our own individual responsibility, these emotions are less likely to influence our decisions.

As with all research, ours comes with important caveats. First, there are other potential explanations for the mixed results observed here, which also draw on the inherent differences between the two emotion-inducing events. We will mention two such possible hypotheses here. First, it could be the case that negative emotions deriving from exam performance have a greater impact on risk tasking because the outcome of the exam has more "consequence" for the individual than the NFL game result. This effect may then operate through the same self-esteem channel discussed earlier. However, this channel is conceptually distinct from the self-responsibility channel, in that it is about the future ramifications of the emotion-inducing events, rather than whether or not that past event was in your control.[10] One could

[10]It should be noted that the explanation of a mediating effect for consequence is somewhat inconsistent with existing work on emotions in lab settings, which often use film clips and other unambiguously "inconsequential" primes to study emotions and decisions. Why would these studies find effects on decisions from such inconsequential manipulations? This is an inconsistency that merits further research.

imagine an experiment that sought to use random or quasi-random assignment to vary this parameter, to further explore this question. For example, a researcher might conduct a laboratory study that uses repeat experimental tasks for cash payments, and manipulate when individuals are given feedback about their poor/good performance. In this way, an emotional response to feedback, and ensuing risk decisions, could either be attributable to past poor performance linked to self-responsibility (if the feedback and low payment came at the end of the study) or to concerns about the consequences of poor performance for long-run prospects in the experimental tasks (if feedback came earlier in the study).

A second possible explanation for the observed mixed results could be that while NFL fans are more accustomed to being disappointed by their team, undergraduates at elite universities who do relatively poorly on a test are not as experienced with disappointing academic outcomes. In other words, there may be either: 1) a role for learning to cope with negative emotion, to prevent it from influencing risk preferences; or 2) a mediating role for one's history of being disappointed by a specific emotion-inducing stimulus. Again, future research could contribute in this area. One could imagine, for example, a quasi-random field experiment where variation in a sports fan's period of fandom could be used as a way to identify "long-suffering" fans and "fresh-faced" fans, to determine if they respond differently to game-related negative emotion. Another, and potentially more ambitious, option would be a longitudinal laboratory study that varied the opportunity to "learn to cope" with negative emotion across subjects, to determine if one can develop a "history" of dealing with negative emotion in a way that prevents such emotion from influencing decisions in other parts of their life.

A second caveat is that our field contexts do not allow us to isolate the effects of individual emotions in a credible way. Therefore, it is difficult to read our results as either supportive or not supportive of some existing theories, such as the Appraisal Tendency Framework (ATF). However, our research does suggest a promising line of inquiry for future research. Namely, researchers could explore the impact of real-world emotional response on risk by focusing on high- versus low-control emotions specifically (for example, anger/happiness versus sadness/anxiety). The ATF predicts that high-control emotions would induce risk-seeking, while low-control emotions would induce risk-aversion. Our argument for a mediating role for self-responsibility in real-world settings might suggest that both of these predictions will be stronger in instances where self-responsibility is higher, and weaker when self-responsibility is lower.

Third, it is important to note that our studies did not vary the nature of the risks themselves in a meaningful way, which existing research suggests could play an important role in explaining any relationship between emotions and risk. For example, Isen and Patrick (1983) suggest that risks need to be sufficiently large in magnitude to influence an individual's affective state, and trigger one's tendency to protect positive affect. Along these lines, Treffers *et al.* (2012) found that sadness (relative to a neutral state) led to more risk aversion only when the financial stakes were absent or low; they found no relationship between induced emotions and risk-taking in high-stake situations. In addition to the magnitude of risks, the source of risks involved in the decision tasks may also influence the link between emotion and risk-taking. For example, Kugler *et al.* (2012) found a preference reversal when risks were generated by the uncertain behavior of another person rather than the randomness of a lottery. Our findings do not speak directly to this literature, but follow-up work could easily incorporate variation in risk magnitude in ways that refine and improve existing literature in this area.

Finally, our study is of course not a controlled, randomized exploration of the topic, by design. However, an alternative approach to exploring the self-responsibility hypothesis we put forward here would be a randomized experiment; for example, one might design a laboratory-based 2x2 study on emotion and risk behavior. Such a study would vary not only the emotional stimulus (inducing sadness versus inducing neutral emotion, for example), but also the extent to which self-responsibility is the source of the emotional stimulus. Again, a prediction based on our results would be that the link between emotion and risk is stronger when the emotion is more closely linked to self-responsibility. Of course, a limitation to generalizing from such a laboratory study would be that our study and other existing work shows that real-world emotions are more complex than emotions induced in the lab (and discrete emotions may cancel each other out in the real world).

In addition to our hypothesis around self-responsibility and its potential mediating role, we draw two main conclusions on the methodological side. First, we believe that our results show that field studies provide a unique and valuable opportunity to observe participants during moments of intense emotion. In our first field study, students who received low grades on their midterm exam reported significantly higher negative feelings than students who did well. This relationship was evident both in terms of absolute score and score relative to expectations. In our second study, we similarly found very large effects of game outcomes on reported emotions. Based on these results, we conclude that the field may be a better setting for studying intense emotions than the laboratory.

Second, it is important to exercise caution in evaluating the results from field studies, and utilizing a formal statistical model is particularly important. For example, in our first field study, we find evidence that performance on the midterm exam is not independent of participants' characteristics, implying that using a cross-sectional instrumental variables approach to interpret our results would be inappropriate. The only way to control for this violation of the identification assumption is to adopt a panel instrumental variables approach with repeat data from each participant. Unfortunately, in our review of previous studies using quasi-experimental designs to study emotions, we find no studies that have used such a design or that have conducted appropriate statistical tests for exogeneity. Similarly, in our second field study, naive regressions found links between risk decisions and emotions that were illusory - they disappeared when an appropriate model—a panel instrumental variable approach—was used for analysis. These results suggest that future researchers would do well to incorporate panel instrumental variable methods into their analytical approach, particularly when studying emotions in the field.

References

ABADIE, A., DRUKKER, D., HERR, J. L. and IMBENS, G. W. (2004). Implementing matching estimators for average treatment effects in stata. *The Stata Journal*, **4** (3), 240–311.

— and IMBENS, G. (2011). *Bias-Corrected Matching Estimators for Average Treatment Effects*. Tech. Rep. 1.

ALLCOTT, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, **95** (9), 1082–1095.

— and MULLAINATHAN, S. (2010). Behavioral science and energy policy. *Science*, **327** (5970), 1204–1205.

— and ROGERS, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, **104** (10), 3003–3037.

— and TAUBINSKY, D. (2014). The lightbulb paradox: Evidence from two randomized experiments. *Working Paper*.

AMERICANWATER (2015). The water leak detection kit. *American Water: Online*.

ANTIN, J. and CHURCHILL, E. (2011). Badges in social media: A social psychological perspective. *CHI 2011 Gamification Workshop Proceedings*.

AQUINO, K. and II, A. R. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, **83** (6).

ARIELY, D. and LOEWENSTEIN, G. (2006). The heat of the moment: The effect of sexual arousal on sexual decision making. *Journal of Behavioral Decision Making*, **19** (2), 87–98.

ARKES, H., HERREN, L. and ISEN, A. (1988). The role of potential loss in the influence of affect on risk-taking behavior. *Organizational Behavior and Human Decision Processes*, **42** (2), 181–193.

ARONSON, E. and MILLS, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology*, **59**, 177–181.

ASHRAF, N., CAMERER, C. and LOEWENSTEIN, G. (2005). Adam smith, behavioral economist. *Journal of Economic Perspectives*, **19** (3), 131–145.

—, Karlan, D. and Yin, W. (2006). Tying odysseus to the mast: Evidence from a commitment savings product in the philippines. *The Quarterly Journal of Economics*, **121** (2), 635–672.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, **84** (2), 191–215.

— (1994). Self-efficacy. In *Encyclopedia of Human Behavior*, New York: Academic Press.

Banerjee, A., Glennerster, R. and Duflo, E. (2008). Putting a band-aid on a corpse: Incentives for nurses in the indian public health care system. *Journal of the European Economic Association*, **6** (2-3), 487–500.

Banerjee, A. V. and Mullainathan, S. (2008). Limited attention and income distribution. *American Economic Review*, **98** (2), 489–493.

Bazelon, E. (2014). How payday lenders prey upon the poor and the courts don't help. *The New York Times Magazine*.

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P. and Kraut, R. E. (2004). Using social psychology to motivate contributions to online communities. *CSCW Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pp. 212–221.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, **119** (1), 249–275.

—, Karlan, D., Mullainathan, S., Shafir, E. and Zinman, J. (2010). What's advertising content worth? evidence from a consumer credit marketing field experiment. *The Quarterly Journal of Economics*, **129** (3), 263–306.

Berument, H. and Ceylan, N. B. (2012). Effects of soccer on stock markets: The return-volatility relationship. *The Social Science Journal*, **49** (3), 368–374.

Beshears, J., Choi, J. J., Laibson, D., Madrian, B. C. and Milkman, K. L. (2015). The effect of providing peer information on retirement savings decisions. *Journal of Finance, Forthcoming*.

Bicchieri, C. and Lev-On, A. (2007). Computer-mediated communication and cooperation in social dilemmas: an experimental analysis. *Politics, Philosophy, and Economics*, **6** (2), 139–168.

Breza, E. (2010). Peer pressure and loan repayment: Evidence from a natural experiment. *Working Paper*.

Brockner, J., Wiesenfeld, B. and Raskas, D. (1983). Self-esteem and expectancy-value discrepancy: The effects of believing that you can (or can't) get what you want. In R. Baumeister (ed.), *Self-Esteem: The Puzzle of Low Self-Regard*, New York: Plenum Press, pp. 219–240.

Brooks, A. and Schweitzer, M. (2011). Can nervous nelly negotiate? how anxiety causes negotiators to make low first offers, exit early, and earn less profit. *Organizational Behavior and Human Decision Processes*, **115** (1), 43–54.

Brown, J. (2014). Governor brown declares drought state of emergency.

— (2015). Governor brown directs first ever statewide mandatory water reductions.

Butler, G. and Mathews, A. (1987). Anticipatory anxiety and risk perception. *Cognitive Therapy and Research*, **11** (5), 551–565.

Cacioppo, J., Gardner, W. and Berntson, G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, **1** (1), 3–25.

Cadena, X. and Schoar, A. (2011). Remembering to pay? reminders vs. financial incentives for loan payments. *NBER Working Paper*, (17020).

Camerer, C. and Loewenstein, G. (2004). Behavioral economics: Past, present, and future. In C. Camerer, G. Loewenstein and M. Rabin (eds.), *Advances in Behavioral Economics*, Princeton: Princeton University Press.

Card, D. and Dahl, G. (2011). Family violence and football: The effect of unexpected emotional cues on violent behavior. *The Quarterly Journal of Economics*, **19** (1), 103–143.

CFPB (2014). Consumer financial protection bureau: Data point - payday lending. *Consumer Financial Protection Bureau: Online*.

CFRL (2014). Center for responsible lending: Fast facts - payday loans. *Center for Responsible Lending: Online*.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, **74** (6), 1579–1601.

Chou, K.-L., Lee, T. and Ho, A. (2007). Does mood state change risk taking tendency in older adults. *Psychology and Aging*, **22** (2), 310–318.

Cialdini, R., Demaine, L., Sagarin, B., Barrett, D., Rhoads, K. and Winter, P. (2006). Managing social norms for persuasive impact. *Social Influence*, **1** (1), 3–15.

Clapp, J., Lange, J., Russell, C., Shillington, A. and Voas, R. (2003). A failed norms social marketing campaign. *Journal of Studies on Alcohol*, **64** (3), 409–414.

Clee, M. and Wicklund, R. (1980). Consumer behavior and psychological reactance. *Journal of Consumer Research*, **6** (4), 389–405.

Cochrane, J. (2010). Interview in pbs nova: Mind over money. [Premiered March 30, 2010].

Cordray, R. (2014). Director richard cordray remarks at the payday field hearing: March 25, 2014. *Consumer Financial Protection Bureau*.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, **47** (2), 448–474.

Cryder, C., Lerner, J., Gross, J. and Dahl, R. (2008). Misery is not miserly. *Psychological Science*, **19** (6), 525–530.

Cunningham, M. (1979). Weather, mood, and helping behavior: Quasi experiments with the sunshine samaritan. *Journal of Personality and Social Psychology*, **37** (11), 1947–1956.

Davis, M. (2011). Behavior and energy savings evidence from a series of experimental interventions. *Environmental Defense Fund Report*.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, **47** (2), 315–372.

—, List, J. and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, **127** (1), 1–56.

DeMartini, K., Leeman, R., Corbin, W. and Toll, B. (2014). A new look at risk-taking: Using a translational approach to examine risk-taking behavior on the balloon analogue risk task. *Experimental and Clinical Psychopharmacology*, **22** (5), 444–452.

Dickerson, C., Thibodeau, R., Aronson, E. and Miller, D. (1992). Using cognitive dissonance to encourage water conservation. *Journal of Applied Social Psychology*, **22** (11), 841–854.

DNI (2012). Ica report: Global water security. *Office of The Director of National Intelligence: February 2, 2012*.

Doherty, D. and Adler, E. S. (2014). The persuasive effects of partisan campaign mailers. *The Political Research Quarterly*, **67** (3), 562–573.

Dustmann, C. and Schonberg, U. (2012). Expansions in maternity leave coverage and children's long-term outcomes. *American Economic Journal: Applied Economics*, **4** (3), 190–224.

Duval, S. and Wicklund, R. A. (1972). *A Theory of Objective Self-Awareness*. Oxford: Academic Press.

Duval, T. S. and Silvia, P. J. (2002). Self-awareness, probability of improvement, and the self-serving bias. *Journal of Personality and Social Psychology*, **82** (1), 49–61.

Eagly, A. and Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist*, **54** (6), 408–423.

Eckel, C. and Grossman, P. (1996). The relative price of fairness: Gender differences in a punishment game. *Journal of Economic Behavior and Organization*, **30** (2), 143–158.

Ellingsen, T. and Johannesson, M. (2004). Promises, threats and fairness. *The Economic Journal*, **114** (495), 397–420.

EPA (2009). Water on tap: What you need to know. *United States Environmental Protection Agency: Office Of Ground Water And Drinking Water*.

Espey, M., Espey, J. and Shaw, W. (1997). Price elasticity of residential demand for water: A meta-analysis. *Water Resources Research*, **33** (6), 1369–1374.

FCA (2014). Fca confirms price cap rules for payday lenders: November 11, 2014. *U.K. Financial Conduct Authority: Press Release*.

Fehr, E. and Gintis, H. (2007). Human motivation and social cooperation: Experimental and analytical foundations. *Annual Review of Sociology*, **33**, 43–64.

Ferraro, P. and Miranda, J. J. (2013). Heterogeneous treatment effects and mechanisms in information-based environmental policies: Evidence from a large-scale field experiment. *Resource and Energy Economics*, **35** (3), 356–379.

—, — and Price, M. (2011). The persistence of treatment effects with norm-based policy instruments: Evidence from a randomized environmental policy experiment. *American Economic Review*, **101** (3), 318–322.

— and Price, M. (2013). Using non-pecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Review of Economics and Statistics*, **95** (1), 64–73.

Fessler, D., Pillsworth, E. and Flamson, T. (2004). Angry men and disgusted women: An evolutionary approach to the influence of emotions on risk taking. *Organizational Behavior and Human Deciison Processes*, **95** (1), 107–123.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, **7** (2), 117–140.

Fischer, C. (2008). Feedback on household electricity consumption: A tool for saving energy? *Energy efficiency*, **1** (1), 79–104.

Forgas, J. (1995). Mood and judgment: The affect infusion model. *Psychological Bulletin*, **117** (1), 39–66.

Fox, J. (2015). From "economic man" to behavioral economics. *The Harvard Business Review*.

Franco, A., Malhotra, N. and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, **345** (6203), 1502–1505.

Garcia, S., Tor, A. and Gonzalez, R. (2006). Ranks and rivals: a theory of competition. *Personality and Social Psychology Bulletin*, **32** (7), 970–82.

Gensler, L. (2014). Cfpb moves against payday loan industry, orders ace cash express to pay 10 million. *Forbes Magazine*.

Gerber, A., Green, D. and Larimer, C. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, **102** (1), 33–48.

Gillingham, K., Newell, R. and Palmer, K. (2009). Energy efficiency economics and policy. *Annual Review of Resource Economics*, **1** (1), 597–620.

Grable, J. and Roszkowski, M. (2008). The influence of mood on the willingness to take financial risks. *Journal of Risk Research*, **11** (7), 905–923.

Granfield, R. (2005). Alcohol use in college: Limitations on the transformation of social norms. *Addiction Research and Theory*, **13** (3), 281–292.

Griskevicius, V., Cialdini, R. B. and Goldstein, N. J. (2008). Social norms: An underestimated and underemployed lever for managing climate change. *International Journal of Sustainability Communication*, **3**, 5–13.

Guéguen, N. (2013). Weather and smiling contagion: A quasi experiment with the smiling sunshine. *Journal of Nonverbal Behavior*, **37** (1), 51–55.

Hagger, M., Chatzisarantis, N. and Biddle, S. (2002). A meta-analytic review of the theories of reasoned action and planned behavior in physical activity: Predictive validity and the contribution of additional variables. *Journal of Sport and Exercise Psychology*, **24** (1), 3–32.

Han, S., Lerner, J. and Zeckhauser, R. (2012). The disgust-promotes-disposal effect. *Journal of Risk and Uncertainty*, **44** (2), 101–113.

Harding, M. and Hsiaw, A. (2014). Goal setting and energy conservation. *Journal of Economic Behavior and Organization*, **107A**, 209–227.

Harrison, G. and List, J. (2004). Field experiments. *Journal of Economic Literature*, **42** (4).

Healy, A., Malhotra, N. and Mo, C. H. (2010). Irrelevant events affect voters' evaluations of government performance. *Proceedings of the National Academy of Sciences*, **107** (29), 12804–12809.

Heilman, R., Crişan, L., Houser, D., Miclea, M. and Miu, A. (2010). Emotion regulation and decision making under risk and uncertainty. *Emotion*, **10** (2), 257–265.

Hirt, E., Zillmann, D., Erickson, G. and Kennedy, C. (1992). Costs and benefits of allegiance: Changes in fans' self-ascribed competencies after team victory versus defeat. *Journal of Personality and Social Psychology*, **63** (5), 724–738.

Isen, A. and Patrick, R. (1983). The effect of positive feelings on risk taking: When the chips are down. *Organizational Behavior and Human Performance*, **31** (2), 194–202.

Jaffe, A. and Stavins, R. (1994). The energy paradox and the diffusion of conservation technology. *Resource and Energy Economics*, **16** (2), 91–122.

James, W. (1890). *Principles of Psychology*. New York: H. Holt and Company.

Jessoe, K. and Rapson, D. (2014). Knowledge is (less) power: Experimental evidence from residential energy use. *American Economic Review*, **104** (4), 1417–1438.

Jevons, W. S. (1871). *The Theory of Political Economy*. London: Macmillan and Co.

John, L. K. and Norton, M. I. (2013). Converging to the lowest common denominator in physical health. *Health Psychology*, **32** (9), 1023–1028.

Jonas, E., Schimel, J., Greenberg, J. and Pyszczynski, T. (2002). The scrooge effect: Evidence that mortality salience increases prosocial attitudes and behavior. *Personality and Social Psychology Bulletin*, **28** (10).

Josephs, R., Larrick, R., Steele, C. and Nisbett, R. (1992). Protecting the self from the negative consequences of risky decisions. *Journal of Personality and Social Psychology*, **62** (1), 26–37.

Julian, J. W. and Perry, F. A. (1967). Cooperation contrasted with intra-group and inter-group competition. *Sociometry*, **30** (1), 79–90.

Kahneman, D. (2011). *Thinking Fast and Slow*. Farrar, Straus and Giroux.

—, Knetsch, J. and Thaler, R. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, **98** (6), 1325–48.

— and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, **47** (2), 263–291.

Karlan, D., McConnell, M., Mullainathan, S. and Zinman, J. (2010). Getting to the top of mind: How reminders increase saving. *NBER Working Paper*, (16205).

—, Morten, M. and Zinman, J. (2015). A personal touch: Text messaging for loan repayment. *Working Paper*.

— and Zinman, J. (2009). Observing unobservables: Identifying information asymmetries with a consumer credit field experiment. *Econometrica*, **77** (6), 1993–2008.

Kast, F., Meier, S. and Pomeranz, D. (2014). Under-savers anonymous: Evidence on self-help groups and peer pressure as a savings commitment device. *Harvard Business School Working Paper*.

Kempton, W., Feuermann, D. and McGarity, A. (1992). I always turn it on "super": User decisions about when and how to operate room air conditioners. *Energy and Buildings*, **18**, 177–191.

— and Montgomery, L. (1982). Folk quantification of energy. *Energy*, **7** (10), 817–827.

Kettle, K. L. and Habul, G. (2011). The signature effect: Signing influences consumption-related behavior by priming self-identity. *Journal of Consumer Research*, **38** (3), 474–489.

Kiel, P. (2013a). How one state succeeded in restricting payday loans. *ProPublica*.

— (2013b). Whack-a-mole: How payday lenders bounce back when states crack down. *ProPublica*.

Kirby, S., Francis, B. and O'Flaherty, R. (2014). Can the fifa world cup football (soccer) tournament be associated with an increase in domestic abuse? *Journal of Research in Crime and Delinquency*, **51** (3), 259–276.

Kohn, A. (1996). *No Contest: The Case Against Competition*, Houghton Mifflin, chap. The "Number One" Obsession.

Kraft-Todd, G., Bhanot, S., Yoeli, E. and Rand, D. (2015). Promoting cooperation in the field. *Current Opinion in Behavioral Sciences*, **3**, 96–101.

Krupić, D. and Corr, P. (2014). Individual differences in emotion elicitation in university examinations: A quasi-experimental study. *Personality and Individual Differences*, **71**, 176–180.

Kugler, T., Connolly, T. and Ordonez, L. (2012). Emotion, decision, and risk: Betting on gambles versus betting on people. *Journal of Behavioral Decision Making*, **25** (2), 123–134.

Kuziemko, I., Buell, R., Reich, T. and Norton, M. I. (2014). "last-place aversion": Evidence and redistributive implications. *The Quarterly Journal of Economics*, **129** (1), 105–149.

Larrick, R., Timmerman, T., Carton, A. and Abrevaya, J. (2011). Temper, temperature, and temptation heat-related retaliation in baseball. *Psychological Science*, **22**, 423–428.

Larsen, J., McGraw, A. P. and Cacioppo, J. (2001). Can people feel happy and sad at the same time? *Journal of Personality and Social Psychology*, **81** (4), 684–696.

Lee, J. and Tanverakul, S. (2015). Price elasticity of residential water demand in california. *Journal of Water Supply*, **64** (2), 211–218.

Leith, K. and Baumeister, R. (1996). Why do bad moods increase self-defeating behavior?: Emotion, risk taking, and self-regulation. *Journal of Personality and Social Psychology*, **71** (6), 1250–1267.

Lejuez, C., Read, J., Kahler, C., Richards, J., Ramsey, S., Stuart, G., Strong, D. and Brown, R. (2002). Evaluation of a behavioral measure of risk taking: The balloon analogue risk task (bart). *Journal of Experimental Psychology: Applied*, **8** (2), 75–84.

Lerner, J. and Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition and Emotion*, **14** (4), 473–493.

— and — (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*, **81** (1), 146–159.

—, Small, D. and Loewenstein, G. (2004). Heart strings and purse strings: Carryover effects of emotions on economic decisions. *Psychological Science*, **15** (5), 337–341.

— and Tiedens, L. (2006). Portrait of the angry decision maker: How appraisal tendencies shape anger's influence on cognition. *Journal of Behavioral Decision Making*, **19** (2), 115–137.

List, J. (2003). Does market experience eliminate market anomalies? *The Quarterly Journal of Economics*, **118** (1), 41–71.

Little, D. (2012). A crucible of competition and cooperation: Where do the concepts fit in recreation activity delivery? *University of Waikato Research Commons*, **10**, 141–154.

Loewenstein, G. and Lerner, J. (2003). The role of affect in decision making. In R. Davidson, K. Sherer and H. H. Goldsmith (eds.), *Handbook of Affective Science*, Oxford: Oxford University Press, pp. 619–642.

Lynne, G., Casey, C. F., Hodges, A. and Rahmani, M. (1995). Conservation technology adoption decisions and the theory of planned behavior. *Journal of Economic Psychology*, **16** (4), 581–598.

MANO, H. (1992). Judgments under distress: Assessing the role of unpleasantness and arousal in judgment formation. *Organizational Behavior and Human Decision Processes*, **52** (2), 216–245.

MAYER, P. and DEOREO, W. (1999). Mean per capita daily water use. In *Residential End Uses of Water*, AWWA Research Foundation and American Water Works Association, p. 87.

MAZAR, N., AMIR, O. and ARIELY, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, **45** (6).

MCCABE, D. and TREVINO, L. K. (1993). Academic dishonesty: Honor codes and other contextual influences. *The Journal of Higher Education*, **43** (3), 522–538.

MILLER, M. (2013). For the win! the effect of professional sports records on mayoral elections. *Social Science Quarterly*, **94** (1), 59–78.

MUTZ, D. and PEMANTLE, R. (2011). The perils of randomization checks in the analysis of experiments. *Paper presented at the Annual Meetings of the Society for Political Methodology*.

NASP (2010). *Self-Efficacy: Helping Children Believe They Can Succeed*. Tech. rep., National Association of School Psychologists.

NORTON, M. I., MOCHON, D. and ARIELY, D. (2011). The "ikea effect": When labor leads to love. *Journal of Consumer Psychology*, **22** (3), 453–460.

OBAMA, B. (2014). Remarks by the president on the california drought: February 14, 2014.

OLIVER, J. (2014). Last week tonight with john oliver: Predatory lending: August 10, 2014.

OLMSTEAD, S. and STAVINS, R. (2007). Managing water demand price vs. non-price conservation programs. *Pioneer Institute for Public Policy Research*, (39).

OLSON, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge: Harvard University Press.

O'NEIL, B. (2011). Review of *Broke, USA: From Pawnshops to Poverty, Inc. How the Working Poor Became Big Business*. *Journal of Financial Counseling and Planning*, **22** (1).

PAJARES, F. (1997). *Advances in Motivation and Achievement*, Greenwich: JAI Press, chap. Current Directions in Self-Efficacy Research.

PECK, J. and CHILDERS, T. (2003). To have and to hold: The influence of haptic information on product judgments. *Journal of Marketing*, **67** (2), 35–48.

— and SHU, S. (2009). The effect of mere touch on perceived ownership. *Journal of Consumer Research*, **36** (3).

PELLIGRA, V. (2011). Empathy, guilt-aversion, and patterns of reciprocity. *Journal of Neuroscience, Psychology, and Economics*, **4** (3), 161–173.

Petersen, J., Shunturov, V., Janda, K., Platt, G. and Weinberger, K. (2007). Dormitory residents reduce electricity consumption when exposed to real-time visual feedback and incentives. *International Journal of Sustainability in Higher Education*, **8** (1), 16–33.

Pierce, L., Rogers, T. and Snyder, J. (2015). Losing hurts: The happiness impact of partisan electoral loss. *Journal of Experimental Political Science (forthcoming)*.

Pleskac, T., Wallsten, T., Wang, P. and Lejuez, C. (2008). Development of an automatic response mode to improve the clinical utility of sequential risk-taking tasks. *Experimental and Clinical Psychopharmacology*, **16** (6), 555–564.

Polivy, J. and Herman, C. P. (1985). Dieting and binging: A causal analysis. *American Psychologist*, **40** (2), 193–201.

Qin, Z., Johnson, D. and Johnson, R. (1995). Cooperative versus competitive efforts and problem solving. *Review of Educational Research*, **65** (2), 129–143.

Raghunathan, R. and Pham, M. (1999). All negative moods are not equal: Motivational influences of anxiety and sadness on decision making. *Organizational Behavior and Human Decision Processes*, **79** (1), 56–77.

Rees, D. and Schnepel, K. (2009). College football games and crime. *Journal of Sports Economics*, **10** (1), 68–87.

Rivlin, G. (2011). *Broke, USA: From Pawnshops to Poverty, Inc. - How the Working Poor Became Big Business*. HarperBusiness.

Rochat, P. (2003). Five levels of self-awareness as they unfold early in life. *Consciousness and Cognition*, **12** (4), 717–731.

Samuelson, W. and Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, **1** (1), 7–59.

Schelling, T. (1960). *The Strategy of Conflict*. Oxford: Oxford University Press.

Scherer, K. (1988). Criteria for emotion-antecedent appraisal: A review. *Cognitive Perspectives on Emotion and Motivation*, **44**, 89–126.

Schneider, M. (2014). Weather, mood, and stock market expectations: When does mood affect investor sentiment? *SSRN Working Paper*.

Schultz, W., Nolan, J., Cialdini, R., Goldstein, N. and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, **18** (5).

Schwarz, N. and Clore, G. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, **45** (3), 513–523.

Shah, A. K., Mullainathan, S. and Shafir, E. (2012). Some consequences of having too little. *Science*, **338** (6107), 682–685.

Shelton, T. and Mahoney, M. (1978). The content and effect of "psyching-up" strategies in weight lifters. *Cognitive Therapy and Research*, **2** (3), 275–284.

Shu, L., Gino, F. and Bazerman, M. (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin*, **37** (3), 330–349.

—, Mazar, N., Gino, F., Ariely, D. and Bazerman, M. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences*, **109** (38), 15197–15200.

Silver-Greenberg, J. (2015). Consumer protection agency seeks limits on payday lenders. *The New York Times: Dealbook*.

Small, D., Lerner, J. and Fischhoff, B. (2006). Emotion priming and attributions for terrorism: Americans' reactions in a national field experiment. *Political Psychology*, **27** (2), 289–298.

Smith, A. (1759). *The Theory of Moral Sentiments*. A. Millar.

Smith, C. and Ellsworth, P. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, **48** (4), 813–838.

— and Lazarus, R. (1993). Appraisal components, core relational themes, and the emotions. *Cognition and Emotion*, **7** (3-4), 233–269.

Soman, D., Ainslie, G., Frederick, S., Li, X., Lynch, J., Moreau, P., Mitchell, A., Read, D., Sawyer, A., Trope, Y., Wertbroch, K. and Zauberman, G. (2005). The psychology of intertemporal discounting: Why are distant events valued differently from proximal ones? *Marketing Letters*, **16** (3-4), 347–360.

Stanovich, K. and West, R. (2000). Individual differences in reasoning: Implications for the rationality debate. *The Behavioral and Brain Sciences*, **23** (5), 645–665.

Stets, J. E. and Burke, P. J. (2000). Identity theory and social identity theory. *Social Psychology Quarterly*, **63** (3), 224–237.

Stone, J. and Cooper, J. (2001). A self-standards model of cognitive dissonance. *Journal of Experimental Social Psychology*, **37** (3).

Subramanian, C. (2013). Nudge back in fashion at white house. *Time Magazine*, [Online; posted August 9, 2013].

Suls, J., Martin, R. and Wheeler, L. (2002). Social comparison: Why, with whom, and with what effect? *Current Directions in Psychological Science*, **11** (5), 159–163.

Tiedens, L. and Linton, S. (2001). Judgment under emotional certainty and uncertainty: The effects of specific emotions on information processing. *Journal of Personality and Social Psychology*, **81** (6), 973–988.

Tran, A. and Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, **96** (9-10), 645–650.

Treffers, T., Koellinger, P. and Picot, A. (2012). In the mood for risk? a random-assignment experiment addressing the effects of moods on risk preferences. *ERIM Report Series Reference*, (ERS-2012-014-ORG).

Tschannen-Moran, M. and Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, **17** (7), 783–805.

Turner, J. (2012). *Symbolic Interactionist Theories of Identity*, Wiley-Blackwell.

Urahn, S. K. (2011). *Payday Lending in America: Who Borrows, Where They Borrow, and Why*. Tech. rep., The Pew Charitable Trusts.

Valdesolo, P. and DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, **17** (6), 476–477.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations. *Econometrica*, **76** (6), 1467–1480.

WaterSense (2015). Fix a leak week. *Environmental Protection Agency*, accessed: April 2015.

Watson, R., Murphy, M., Kilfoyle, F. and Moore, S. (1999). An opportunistic experiment in community water conservation. *Population and Environment*, **20** (6), 545–560.

Wechsler, H., Nelson, T. F., Lee, J. E., Seibring, M., Lewis, C. and Keeling, R. (2003). Perception and reality: A national evaluation of social norms marketing interventions to reduce college students' heavy alcohol use. *Journal of Studies on Alcohol*.

Winter, P. (2008). Science notes: Park signs and visitor behavior: A research summary. *Park Science*, pp. 34–35.

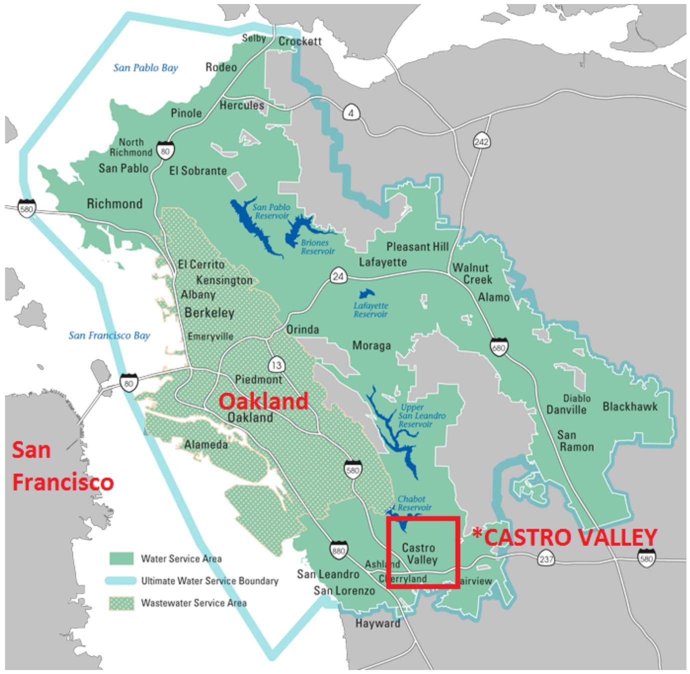
Yim, J. and Graham, T. N. (2007). Using games to increase exercise motivation. *Proceedings of the 2007 Conference on Future Play*, pp. 166–173.

Zeckhauser, R. (1986). Comments: Behavioral versus rational economics: What you see is what you conquer. *Journal of Business*, **59** (4), S435–S449.

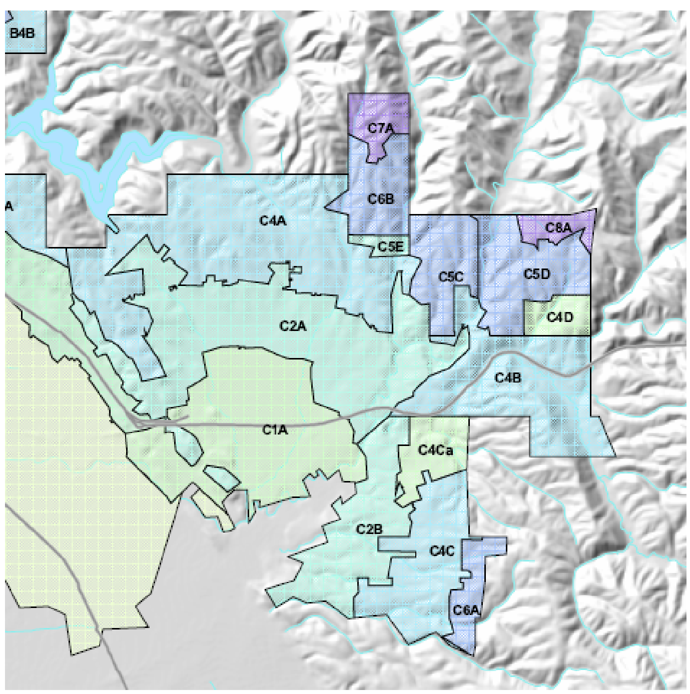
Appendix A

Appendix to Chapter 1

Appendix A.1

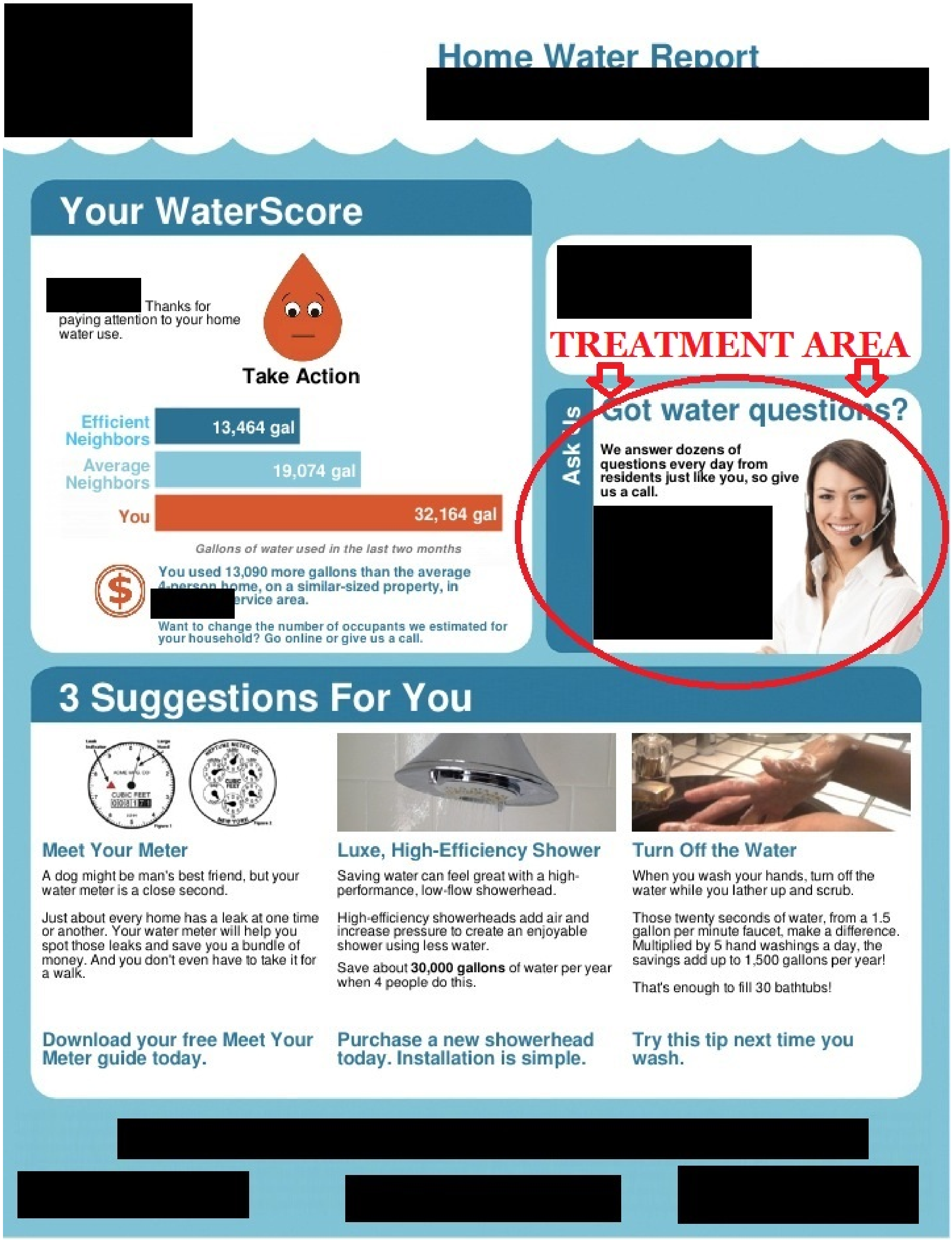


Panel (A.1.1): *Experiment Location*



Panel (A.1.2): *Pressure Zones*

Appendix A.2



[redacted]

Home Water Report

[redacted]

Your WaterScore

[redacted] Thanks for paying attention to your home water use.

Take Action

Efficient Neighbors	13,464 gal
Average Neighbors	19,074 gal
You	32,164 gal

Gallons of water used in the last two months

You used 13,090 more gallons than the average 4-person home, on a similar-sized property, in [redacted]ervice area.

Want to change the number of occupants we estimated for your household? Go online or give us a call.

[redacted]

TREATMENT AREA

Ask Us

Got water questions?

We answer dozens of questions every day from residents just like you, so give us a call.

[redacted]

3 Suggestions For You

Meet Your Meter

A dog might be man's best friend, but your water meter is a close second.

Just about every home has a leak at one time or another. Your water meter will help you spot those leaks and save you a bundle of money. And you don't even have to take it for a walk.

Download your free Meet Your Meter guide today.

Luxe, High-Efficiency Shower

Saving water can feel great with a high-performance, low-flow showerhead.

High-efficiency showerheads add air and increase pressure to create an enjoyable shower using less water.

Save about **30,000 gallons** of water per year when 4 people do this.

Purchase a new showerhead today. Installation is simple.

Turn Off the Water

When you wash your hands, turn off the water while you lather up and scrub.

Those twenty seconds of water, from a 1.5 gallon per minute faucet, make a difference. Multiplied by 5 hand washings a day, the savings add up to 1,500 gallons per year!

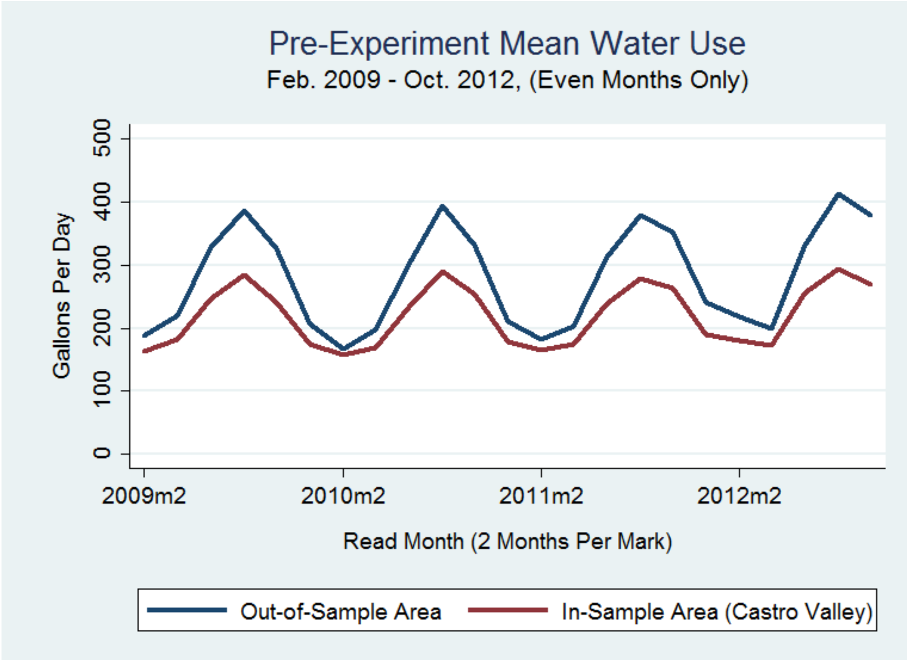
That's enough to fill 30 bathtubs!

Try this tip next time you wash.

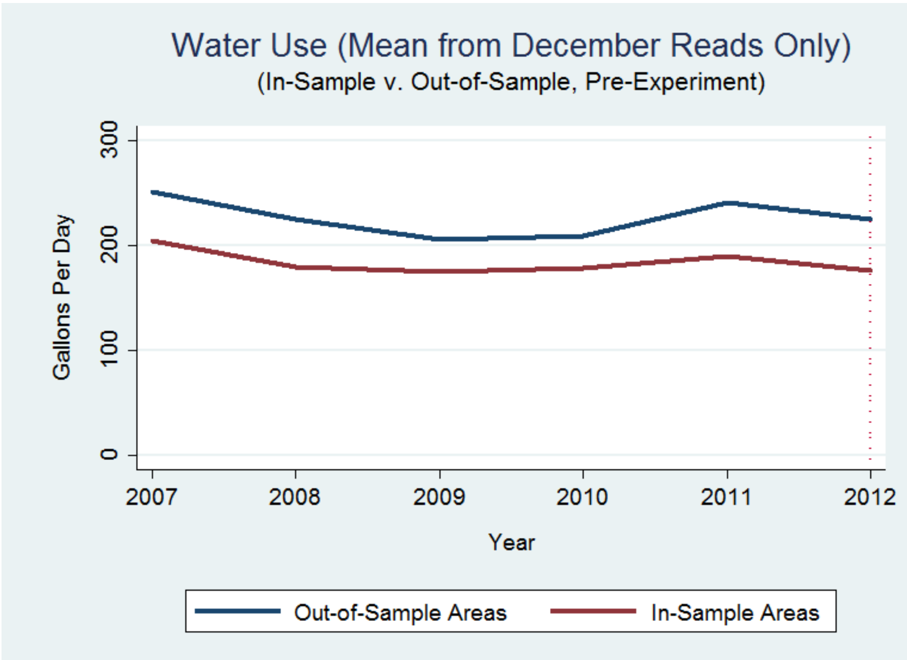
[redacted]

Panel (A.2.1): *Home Water Report*

Appendix A.3

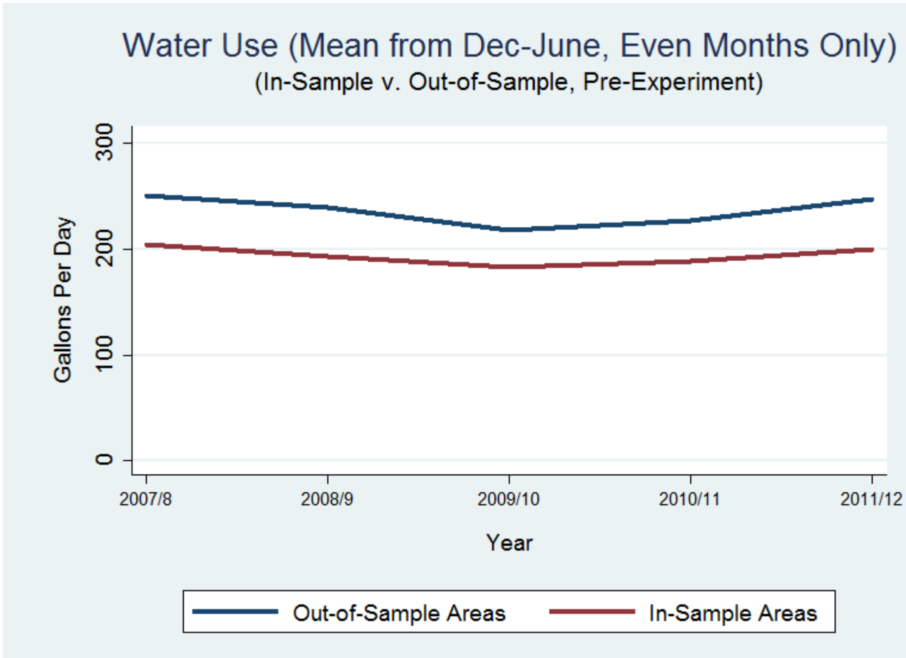


Panel (A.3.1): *Pre-Experiment Water Use (Even Months)*



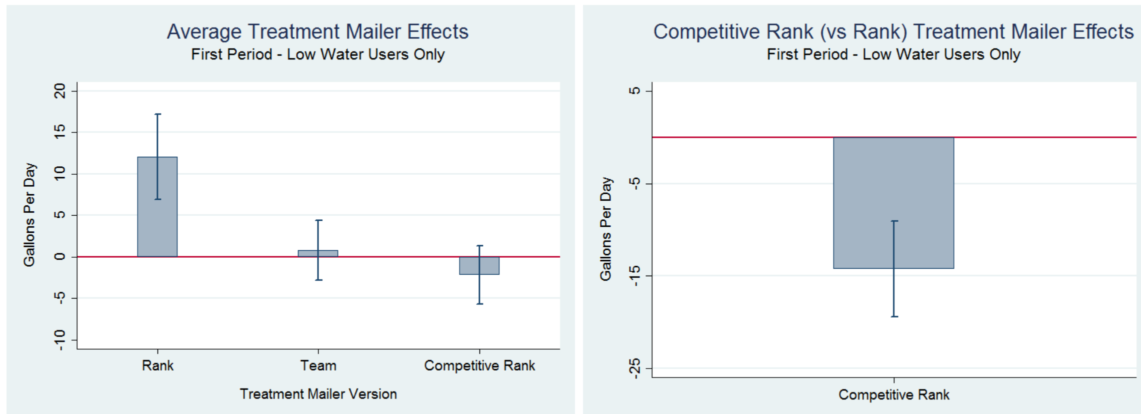
Panel (A.3.2): *Pre-Experiment Water Use (December Only)*

Appendix A.3 (continued)

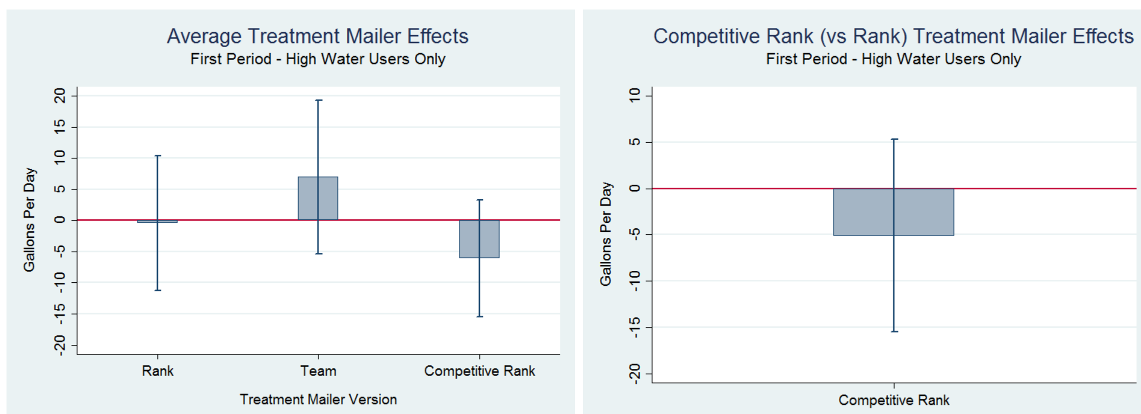


Panel (A.3.3): *Pre-Experiment Water Use (Annual Means)*

Appendix A.4: Short-Run ATE by Prior Water Use (Dec 2012 Only)



Panel (A.4.1): *Short-Run ATE (Low Water Users Only) - SE Marked*



Panel (A.4.2): *Short-Run ATE (High Water Users Only) - SE Marked*

Table A.1: *Demographic Variables by Treatment Group*

	(1) Control Mailer	(2) Rank Mailer	(3) Team Mailer	(4) Comp. Rank Mailer	(5) OOS Control
Home Size (SqFt)	1650.5	1627.5	1628.0	1622.2	1976.2
	(561.7)	(533.4)	(544.0)	(538.3)	(932.2)
Lot Size (SqFt)	7503.7	7320.8	7657.0	7376.4	8655.2
	(4390.9)	(3615.7)	(4821.9)	(5175.4)	(9397.3)
Year Home Built	1958.0	1957.5	1957.4	1957.4	1954.3
	(13.81)	(13.46)	(13.24)	(13.05)	(27.71)
Number of Bedrooms	3.129	3.110	3.160	3.109	3.181
	(0.725)	(0.735)	(0.759)	(0.757)	(0.968)
Number of Bathrooms	2.071	2.055	2.049	2.022	2.311
	(0.827)	(0.859)	(0.832)	(0.833)	(1.066)
Mean Water Use (2012)	229.7	231.0	230.5	236.1	304.0
	(129.6)	(136.9)	(141.3)	(136.9)	(264.5)
N	1308	1288	1284	1300	2880

Means, with standard deviations in parentheses.

Columns (1)-(4) present demographics for the households receiving mailers in this experiment, by mailer version.

Column (5) presents demographics for the out-of-sample control group (a neighboring town not receiving mailers in this experiment).

Table A.2: *Randomization Checks*

	(1) Home Size	(2) Lot Size	(3) Year Built	(4) Bedrooms	(5) Bathrooms	(6) Pre-Treat Mean Water Use
Rank	-25.26	-203.9	-0.565	-0.0185	-0.0147	0.859
	(23.04)	(169.4)	(0.575)	(0.0309)	(0.0354)	(5.264)
CompetitiveRank	-29.98	-120.7	-0.651	-0.0172	-0.0475	5.912
	(23.09)	(202.6)	(0.564)	(0.0312)	(0.0346)	(5.253)
Team	-19.68	163.6	-0.564	0.0352	-0.0135	0.853
	(23.36)	(196.3)	(0.572)	(0.0315)	(0.0349)	(5.374)
Observations	4485	4485	4485	4445	4518	5090
R^2	0.000	0.001	0.000	0.001	0.000	0.000

Standard errors in parentheses.

This table presents the regressions of various household characteristics on dummies for the three treatment groups (Rank, Competitive Rank, and Team), as a randomization check using in-sample data.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.3: *Difference-in-Differences: Dec 2011 (Pre-Treat) vs. Dec 2012 (Post-Treat)*

	(1) GPD	(2) Log GPD
Control*Post	1.650	0.00808
	(8.539)	(0.0355)
Rank*Post	4.237	0.0136
	(8.567)	(0.0357)
Team*Post	4.261	0.00241
	(8.823)	(0.0352)
CompRank*Post	2.107	0.00407
	(8.514)	(0.0353)
Control	-50.20***	-0.207***
	(6.004)	(0.0251)
Rank	-52.49***	-0.216***
	(5.940)	(0.0251)
Team	-51.70***	-0.204***
	(5.873)	(0.0247)
CompRank	-48.53***	-0.190***
	(5.907)	(0.0249)
Post	-17.41**	-0.0887***
	(6.970)	(0.0241)
N	12658	12658
R^2	0.024	0.022

Standard errors in parentheses.
This table presents difference-in-differences regression results comparing the water use of the In-Sample and Out-of-Sample groups. The pre-period is the Dec 2011 read and the post-period is the Dec 2012 read, the period after the first experimental mailer was sent. Regressions without controls are presented, in both level-level and log-level forms.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.4: *Difference-in-Differences: Pre-Treat Mean vs. Post-Treat Mean*

	(1) Mean GPD	(2) Log Mean GPD
Control*Post	-16.17**	-0.0336
	(7.958)	(0.0324)
Rank*Post	-16.28**	-0.0335
	(7.939)	(0.0323)
Team*Post	-14.62*	-0.0459
	(8.539)	(0.0320)
CompRank*Post	-13.44*	-0.0323
	(8.103)	(0.0325)
Control	-46.55***	-0.172***
	(5.348)	(0.0229)
Rank	-47.32***	-0.175***
	(5.355)	(0.0228)
Team	-47.08***	-0.166***
	(5.378)	(0.0224)
CompRank	-43.45***	-0.157***
	(5.397)	(0.0227)
Post	14.58**	0.0369*
	(6.489)	(0.0223)
N	13164	13164
R^2	0.027	0.018

Standard errors in parentheses.

This table presents difference-in-differences regression results comparing the water use of the In-Sample and Out-of-Sample groups. The pre-period water use measure is the mean household water use in the Dec 2011, Feb 2012, Apr 2012, and Jun 2012 reads and the post-period is the mean household water use in the Dec 2012, Feb 2013, Apr 2013, and Jun 2013 reads. Regressions without controls are presented, in both level-level and log-level forms.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.5: *Means Comparison: Water Use in the First Post-Mailer Period (In-Sample Only)*

	(1) GPD	(2) GPD	(3) GPD	(4) Log GPD	(5) Log GPD	(6) Log GPD
Rank	0.774	5.270	4.598	0.000105	0.0279	0.0227
	(4.767)	(5.027)	(4.438)	(0.0256)	(0.0267)	(0.0227)
Team	0.427	2.743	4.306	-0.00547	0.00991	0.0198
	(5.199)	(5.448)	(4.893)	(0.0255)	(0.0268)	(0.0227)
CompetitiveRank	1.751	2.148	-2.051	0.00595	0.00732	-0.0123
	(4.719)	(4.825)	(4.109)	(0.0255)	(0.0267)	(0.0223)
Lot Size		0.00427***	0.00489***		0.0000111***	0.0000156***
		(0.000874)	(0.000717)		(0.00000248)	(0.00000230)
Num Bathrooms		5.512*	2.576		0.0581***	0.0385**
		(3.109)	(2.676)		(0.0184)	(0.0151)
Home Size (SqFt)		0.0269***	0.0229***		0.000120***	0.0000979***
		(0.00562)	(0.00477)		(0.0000303)	(0.0000247)
Constant	175.5***	86.97***	74.15***	4.980***	4.572***	4.553***
	(3.266)	(8.422)	(7.262)	(0.0180)	(0.0366)	(0.0320)
Observations	5041	4440	4414	5039	4439	4414
R^2	0.000	0.054	0.237	0.000	0.038	0.306
Read Month Fixed Effects	No	No	Yes	No	No	Yes
WaterScore Fixed Effects	No	No	Yes	No	No	Yes

Standard errors in parentheses.

This table first presents three regressions (1-3) that compare water use in the first period following mailer initiation in the in-sample groups only. The next three regressions (4-6) present the same results using a log-level specification. The controls used are for household characteristics, the WaterScore that households observed on the first mailer, and Read Month.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.6: *Means Comparison: Water Use in All Periods (In-Sample Only)*

	(1) Mean GPD	(2) Mean GPD	(3) Log Mean GPD	(4) Log Mean GPD
Rank	-0.740	2.398	-0.00413	0.0172
	(4.356)	(4.427)	(0.0227)	(0.0234)
Team	1.391	2.894	-0.00426	0.00799
	(5.063)	(5.105)	(0.0228)	(0.0237)
CompetitiveRank	6.194	8.194*	0.0107	0.0189
	(4.616)	(4.663)	(0.0234)	(0.0243)
Lot Size		0.00628***		0.0000150***
		(0.00163)		(0.00000254)
Num Bathrooms		7.061**		0.0487***
		(3.127)		(0.0156)
Home Size (SqFt)		0.0285***		0.000135***
		(0.00541)		(0.0000253)
Observations	5041	4440	5039	4439
R^2	0.000	0.100	0.000	0.057

Standard errors in parentheses.

This table first presents a simple means comparison of mean water use in all periods following mailer initiation in the in-sample groups only. The subsequent regression controls for household characteristics. The next two regression repeat the analysis, using a log-level specification.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.7: *Low Water Users in the First Post-Mailer Period (In-Sample Only)*

	(1) GPD	(2) GPD	(3) GPD	(4) Log GPD	(5) Log GPD	(6) Log GPD
Rank	9.213**	12.18**	12.02**	0.0485	0.0776*	0.0780*
	(4.655)	(5.139)	(5.165)	(0.0404)	(0.0428)	(0.0427)
Team	-0.423	-0.0609	0.815	0.00199	0.00479	0.0168
	(3.391)	(3.607)	(3.612)	(0.0384)	(0.0411)	(0.0408)
CompetitiveRank	-1.730	-2.461	-2.208	-0.0251	-0.0339	-0.0251
	(3.437)	(3.528)	(3.534)	(0.0389)	(0.0411)	(0.0408)
Lot Size		0.000740	0.00104*		0.00000494	0.00000710
		(0.000616)	(0.000619)		(0.00000497)	(0.00000499)
Num Bathrooms		4.083*	4.027**		0.0546*	0.0543**
		(2.091)	(2.044)		(0.0280)	(0.0274)
Home Size (SqFt)		0.00105	0.00129		0.00000373	0.00000865
		(0.00365)	(0.00361)		(0.0000500)	(0.0000492)
Constant	95.42***	80.75***	86.62***	4.425***	4.279***	4.364***
	(2.418)	(5.537)	(5.654)	(0.0274)	(0.0600)	(0.0595)
Observations	1651	1446	1434	1649	1445	1434
R^2	0.005	0.015	0.036	0.002	0.013	0.034
Read Month Fixed Effects	No	No	Yes	No	No	Yes
WaterScore Fixed Effects	No	No	Yes	No	No	Yes

Standard errors in parentheses.

This table first presents three level regressions (1-3), with different controls, providing means comparisons of water use in the first period following mailer initiation amongst households who were low water users in advance of the experiment. This means that they were in the lowest third of water users in their irrigable area category in the 2012 months that preceded the experiment. The next three regressions (4-6) provide the same results in log-level form. The controls used are for household characteristics, the WaterScore that households observed on the mailer, and the meter read month.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.8: *Low Water Users in the First Post-Mailer Period (Rank and CompRank Only)*

	(1) GPD	(2) GPD	(3) Log GPD	(4) Log GPD
CompetitiveRank	-14.64***	-14.22***	-0.111**	-0.103**
	(5.216)	(5.211)	(0.0436)	(0.0435)
Lot Size	0.000329	0.000643	0.00000291	0.00000532
	(0.000592)	(0.000588)	(0.00000607)	(0.00000603)
Num Bathrooms	1.834	1.941	0.0295	0.0321
	(2.583)	(2.548)	(0.0321)	(0.0315)
Home Size (SqFt)	0.00625	0.00606	0.0000892*	0.0000855*
	(0.00504)	(0.00502)	(0.0000522)	(0.0000516)
Constant	92.22***	95.13***	4.289***	4.352***
	(9.324)	(9.002)	(0.0849)	(0.0850)
Observations	716	708	716	708
R^2	0.015	0.044	0.020	0.036
Read Month Fixed Effects	No	Yes	No	Yes
WaterScore Fixed Effects	No	Yes	No	Yes

Standard errors in parentheses.

This table first presents two regressions (1-2) providing a simple means comparison of water use in the first period following mailer initiation amongst households who were low water users in advance of the experiment. This means that they were in the lowest third of water users in their irrigable area category in the 2012 months that preceded the experiment. The subsequent regressions (3-4) provide the same result in log-level form. The controls used are for household characteristics, the WaterScore that households observed on the mailer, and the meter read month.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.9: *Low Water Users in All Post-Mailer Periods (In-Sample Only)*

	(1)	(2)	(3)	(4)
	Mean GPD	Mean GPD	Log Mean GPD	Log Mean GPD
Rank	2.166	3.850	0.0110	0.0298
	(3.613)	(3.925)	(0.0335)	(0.0357)
Team	-2.658	-2.532	-0.0150	-0.0121
	(3.286)	(3.526)	(0.0325)	(0.0349)
CompetitiveRank	-2.974	-3.255	-0.0362	-0.0400
	(3.400)	(3.604)	(0.0340)	(0.0362)
Lot Size		0.000834^{**}		0.00000676^{**}
		(0.000358)		(0.00000343)
Num Bathrooms		3.402^{*}		0.0338
		(1.932)		(0.0210)
Home Size (SqFt)		0.00194		0.0000279
		(0.00327)		(0.0000364)
Observations	1651	1446	1649	1445
R^2	0.002	0.013	0.001	0.012

Standard errors in parentheses.

This table first presents two level regressions (1-2) of water use in all periods following mailer initiation on the mailer treatments for households who were low water users in advance of the experiment (this means that they were in the lowest third of water users in their irrigable area category in the 2012 months that preceded the experiment. The two regressions that follow (3-4) provide the same results in log-level form. Regressions 2 and 4 use demographic controls.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.10: *Low Water Users in All Post-Mailer Periods (Rank and Competitive Rank Only)*

	(1) Mean GPD	(2) Mean GPD	(3) Log Mean GPD	(4) Log Mean GPD
CompetitiveRank	-5.140	-7.150*	-0.0472	-0.0697*
	(3.536)	(3.913)	(0.0343)	(0.0373)
Lot Size		0.000719		0.00000489
		(0.000571)		(0.00000551)
Num Bathrooms		0.943		0.0199
		(2.387)		(0.0255)
Home Size (SqFt)		0.00600		0.0000793*
		(0.00465)		(0.0000439)
Observations	827	716	827	716
R^2	0.003	0.012	0.002	0.016

Standard errors in parentheses.
This table presents both log (1-2) and level (3-4) regressions of mean water use in all periods following mailer initiation amongst households who were low water users in advance of the experiment (this means that they were in the lowest third of water users in their irrigable area category in the 2012 months that preceded the experiment.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.11: *High Water Users in the First Post-Mailer Period (In-Sample Only)*

	(1) GPD	(2) GPD	(3) GPD	(4) Log GPD	(5) Log GPD	(6) Log GPD
Rank	-5.248	-2.275	-0.453	-0.0269	-0.0175	-0.0122
	(10.11)	(11.04)	(10.83)	(0.0327)	(0.0347)	(0.0340)
Team	1.731	4.287	6.950	-0.0105	0.000372	0.00947
	(12.19)	(12.80)	(12.44)	(0.0324)	(0.0337)	(0.0332)
CompetitiveRank	-4.986	-4.691	-6.061	-0.0158	-0.0134	-0.0130
	(9.650)	(10.01)	(9.432)	(0.0306)	(0.0323)	(0.0316)
Lot Size		0.00637***	0.00611***		0.0000127***	0.0000122***
		(0.00115)	(0.00105)		(0.00000222)	(0.00000213)
Num Bathrooms		-10.41	-8.805		-0.0187	-0.0138
		(6.762)	(6.743)		(0.0202)	(0.0200)
Home Size (SqFt)		0.0306***	0.0319***		0.0000927***	0.000102***
		(0.0108)	(0.0104)		(0.0000305)	(0.0000291)
Constant	267.8***	187.6***	147.0***	5.482***	5.262***	5.105***
	(6.845)	(14.41)	(21.10)	(0.0224)	(0.0422)	(0.0505)
Observations	1692	1459	1454	1692	1459	1454
R^2	0.000	0.067	0.078	0.000	0.045	0.068
Read Month Fixed Effects	No	No	Yes	No	No	Yes
WaterScore Fixed Effects	No	No	Yes	No	No	Yes

Standard errors in parentheses.

This table first presents three level regressions (1-3), with different controls, providing means comparisons of water use in the first period following mailer initiation amongst households who were high water users in advance of the experiment. This means that they were in the highest third of water users in their irrigable area category in the 2012 months that preceded the experiment. The next three regressions (4-6) provide the same results in log-level form. The controls used are for household characteristics, the WaterScore that households observed on the mailer, and the meter read month.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.12: *High Water Users in the First Post-Mailer Period (Rank and CompRank Only)*

	(1) GPD	(2) GPD	(3) Log GPD	(4) Log GPD
CompetitiveRank	-2.709	-5.111	0.00350	-0.000215
	(10.78)	(10.45)	(0.0335)	(0.0330)
Lot Size	0.00502***	0.00501***	0.00000963***	0.00000956***
	(0.00100)	(0.000985)	(0.00000194)	(0.00000200)
Num Bathrooms	-17.46	-16.06	-0.0318	-0.0328
	(10.68)	(10.02)	(0.0283)	(0.0273)
Home Size (SqFt)	0.0479***	0.0467***	0.000127***	0.000136***
	(0.0179)	(0.0165)	(0.0000452)	(0.0000412)
Constant	181.3***	139.3***	5.238***	5.127***
	(17.24)	(19.64)	(0.0534)	(0.0600)
Observations	746	745	746	745
R^2	0.079	0.093	0.044	0.063
Read Month Fixed Effects	No	Yes	No	Yes
WaterScore Fixed Effects	No	Yes	No	Yes

Standard errors in parentheses.

This table first presents two regressions (1-2) providing a simple means comparison of water use in the first period following mailer initiation amongst households who were high water users in advance of the experiment. This means that they were in the highest third of water users in their irrigable area category in the 2012 months that preceded the experiment. The subsequent regressions (3-4) provide the same result in log-level form. The controls used are for household characteristics, the WaterScore that households observed on the mailer, and the meter read month.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.13: *High Water Users in All Post-Mailer Periods (In-Sample Only)*

	(1)	(2)	(3)	(4)
	Mean GPD	Mean GPD	Log Mean GPD	Log Mean GPD
Rank	-1.126	-2.999	-0.00473	-0.00565
	(8.211)	(8.447)	(0.0251)	(0.0255)
Team	11.31	11.69	0.0196	0.0203
	(11.25)	(10.78)	(0.0256)	(0.0261)
CompetitiveRank	9.934	13.28	0.0224	0.0344
	(8.513)	(8.479)	(0.0254)	(0.0258)
Lot Size		0.00955***		0.0000171***
		(0.00309)		(0.00000383)
Num Bathrooms		-8.389		-0.0277*
		(5.749)		(0.0149)
Home Size (SqFt)		0.0290***		0.000102***
		(0.00972)		(0.0000235)
Observations	1692	1459	1692	1459
R^2	0.002	0.172	0.001	0.108

Standard errors in parentheses.

This table first presents two level regressions (1-2) of water use in all periods following mailer initiation on the mailer treatments for households who were high water users in advance of the experiment (this means that they were in the highest third of water users in their irrigable area category in the 2012 months that preceded the experiment. The two regressions that follow (3-4) provide the same results in log-level form. Regressions 2 and 4 use demographic controls.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.14: *High Water Users in All Post-Mailer Periods (Rank and Competitive Rank Only)*

	(1) Mean GPD	(2) Mean GPD	(3) Log Mean GPD	(4) Log Mean GPD
CompetitiveRank	11.06	15.89*	0.0271	0.0399
	(8.455)	(8.310)	(0.0256)	(0.0260)
Lot Size		0.00650***		0.0000140***
		(0.000914)		(0.00000378)
Num Bathrooms		-14.76*		-0.0338
		(7.967)		(0.0218)
Home Size (SqFt)		0.0490***		0.000137***
		(0.0128)		(0.0000347)
Observations	875	746	875	746
R^2	0.002	0.177	0.001	0.112

Standard errors in parentheses.
This table presents both log and level regressions of mean water use in all periods following mailer initiation amongst households who were high water users in advance of the experiment (this means that they were in the highest third of water users in their irrigable area category in the 2012 months that preceded the experiment.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.15: *Last Place Effect: Using All Post-Treatment Periods*

	(1) GPD	(2) GPD	(3) GPD	(4) GPD
Rank	9.297	2.552	1.014	1.347
	(10.48)	(7.385)	(6.009)	(6.087)
Competitive Rank	29.88***	21.00***	18.94***	17.85***
	(10.92)	(7.441)	(5.957)	(5.957)
Mailer GPD		0.448***	0.608***	0.588***
		(0.0444)	(0.0457)	(0.0509)
Lot Size		0.00305***	0.00211***	0.00226***
		(0.000570)	(0.000525)	(0.000540)
Num Bathrooms		-2.511	1.888	1.623
		(5.594)	(4.729)	(4.664)
Home Size (SqFt)		0.0269***	0.0156**	0.0168**
		(0.00799)	(0.00669)	(0.00662)
Constant	286.7***	77.84***	-71.64***	-86.57***
	(7.634)	(12.23)	(16.41)	(13.87)
Observations	3161	2731	2731	2727
R^2	0.005	0.298	0.494	0.497
Read Month Fixed Effects	No	No	Yes	Yes
Mailers Seen Fixed Effects	No	No	Yes	Yes
WaterScore Fixed Effects	No	No	No	Yes

Clustered standard errors, at the household level, in parentheses.
This table presents regression results comparing the effect of the Neutral and Competitive treatments for 'last place' households on subsequent water use. The omitted group here is households in the In-Sample Control who 'would have' been in last place in their groups had they seen a ranking in their mailer.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.16: *First Place Effect: Using All Post-Treatment Periods*

	(1) GPD	(2) GPD	(3) GPD	(4) GPD
Rank	-10.16**	1.269	1.154	2.578
	(4.670)	(3.609)	(3.340)	(3.391)
Competitive Rank	-12.69***	0.0177	-0.0711	1.008
	(4.597)	(3.645)	(3.463)	(3.485)
Mailer GPD		0.731***	0.810***	0.892***
		(0.0366)	(0.0354)	(0.0419)
Lot Size		0.00116**	0.00105**	0.00101**
		(0.000504)	(0.000457)	(0.000452)
Num Bathrooms		1.350	2.288	2.207
		(2.450)	(2.244)	(2.214)
Home Size (SqFt)		0.0104**	0.00878*	0.00746
		(0.00502)	(0.00466)	(0.00459)
Constant	128.1***	21.85***	-20.76***	-36.83***
	(3.064)	(7.359)	(7.793)	(8.675)
Observations	2948	2649	2649	2649
R^2	0.004	0.217	0.341	0.344
Read Month Fixed Effects	No	No	Yes	Yes
Mailers Seen Fixed Effects	No	No	Yes	Yes
WaterScore Fixed Effects	No	No	No	Yes

Clustered standard errors, at the household level, in parentheses.
This table presents regression results comparing the effect of the Neutral and Competitive treatments for 'first place' households on subsequent water use. The omitted group here is households in the In-Sample Control who 'would have' been in first place in their groups had they seen a ranking in their mailer.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.17: *Differential Impact of Rank by Treatment (Control 3rd Place Omitted)*

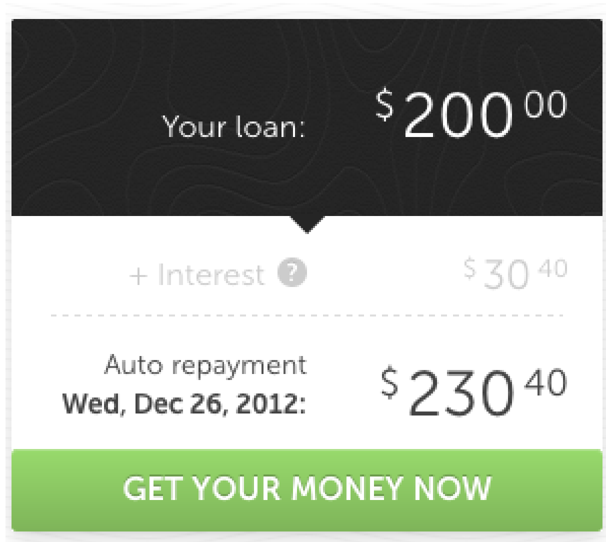
	(1) GPD	(2) GPD	(3) GPD
Control 1st Place	-7.053	-0.920	0.988
	(6.292)	(5.685)	(5.848)
Control 2nd Place	-3.765	-5.892	-5.542
	(5.117)	(4.750)	(4.788)
Control 4th Place	2.543	2.158	2.542
	(5.941)	(5.379)	(5.393)
Control 5th Place	-3.339	-7.189	-6.488
	(7.787)	(7.076)	(7.098)
Rank 1st Place	-8.378	-4.129	-2.587
	(6.549)	(5.934)	(6.034)
Rank 2nd Place	-10.47**	-6.774	-6.156
	(5.078)	(4.565)	(4.590)
Rank 3rd Place	-7.628	-5.335	-5.090
	(5.175)	(4.748)	(4.775)
Rank 4th Place	-2.882	-5.641	-5.209
	(5.450)	(4.853)	(4.890)
Rank 5th Place	12.09	9.293	10.00
	(8.202)	(7.472)	(7.511)
Comp. Rank 1st Place	-10.25*	-7.027	-5.194
	(6.044)	(5.682)	(5.842)
Comp. Rank 2nd Place	-9.333*	-7.082	-6.495
	(4.996)	(4.639)	(4.666)
Comp. Rank 3rd Place	-10.55**	-8.157*	-7.878*
	(5.131)	(4.751)	(4.767)
Comp. Rank 4th Place	-3.106	-4.824	-4.464
	(5.929)	(5.323)	(5.359)
Comp. Rank 5th Place	10.90	7.997	8.618
	(8.828)	(7.852)	(7.917)
Constant	134.4***	34.77***	36.30***
	(5.724)	(7.502)	(7.719)
Observations	4553	4121	4121
R^2	0.101	0.398	0.398
Read Month Fixed Effects	No	Yes	Yes
Demographic and Mailer Number Fixed Effects	No	Yes	Yes
WaterScore Fixed Effects	No	No	Yes

Regressions omit Control 3rd Place households. Clustered standard errors in parentheses.
Regressions include only the middle third of water users, based on use pre-experiment in 2012.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

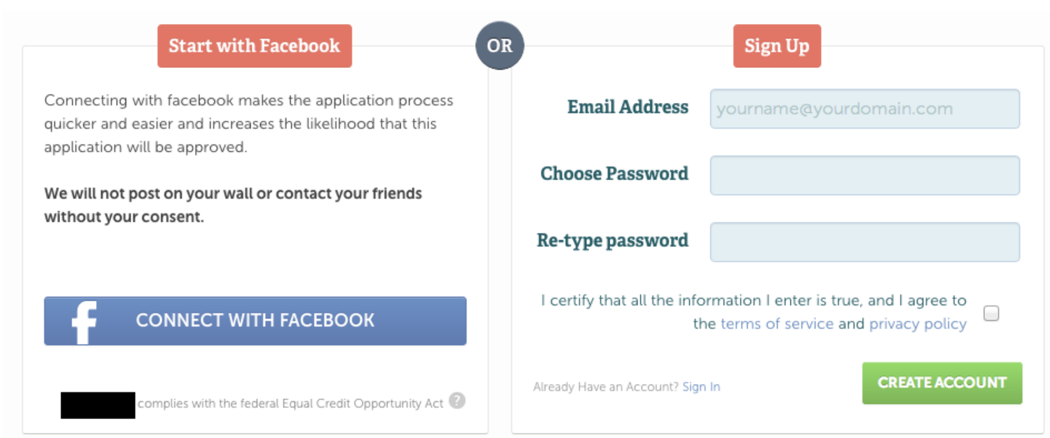
Appendix B

Appendix to Chapter 2

Appendix B.1: Sign Up Screens

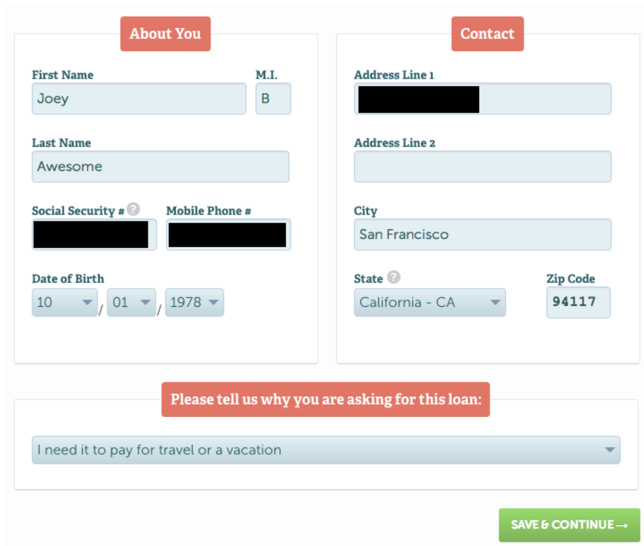


Panel (B.1.1): *Screen #1: Initiate Process*

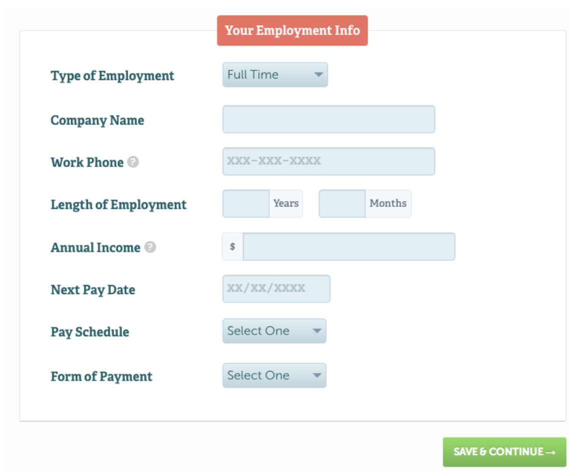


Panel (B.1.2): *Screen #2: Sign Up*

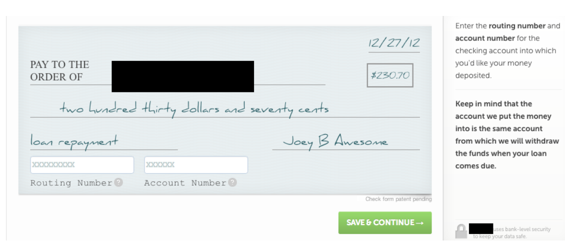
Appendix B.1: Sign Up Screens (continued)



Panel (B.1.3): *Screen #3: Personal Info*

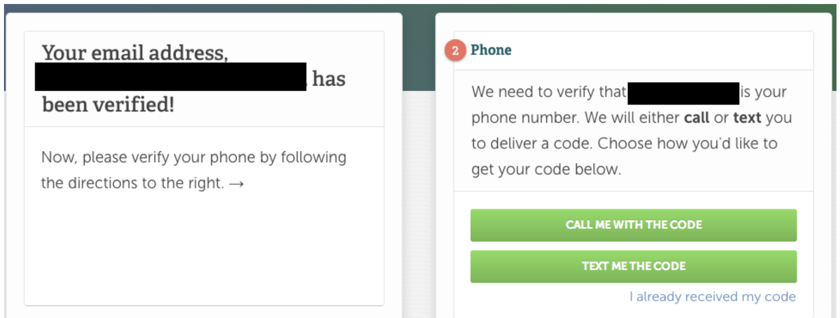


Panel (B.1.4): *Screen #4: Employment Info*

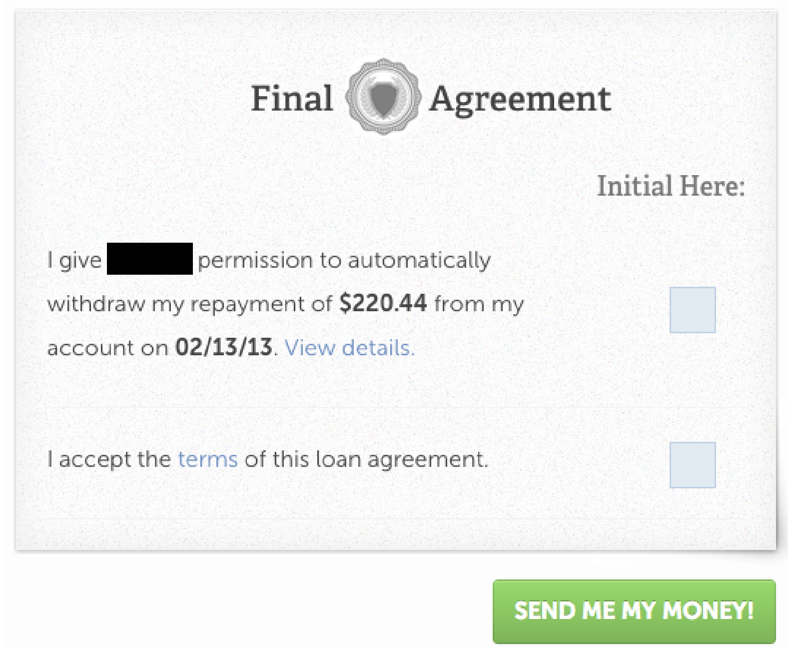


Panel (B.1.5): *Screen #5: Direct Deposit Information*

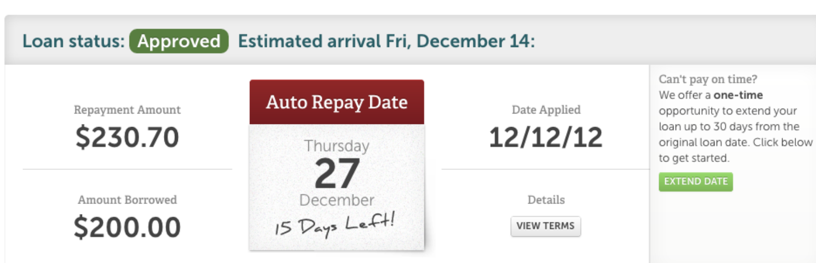
Appendix B.1: Sign Up Screens (continued)



Panel (B.1.6): *Screen #6: Verification*

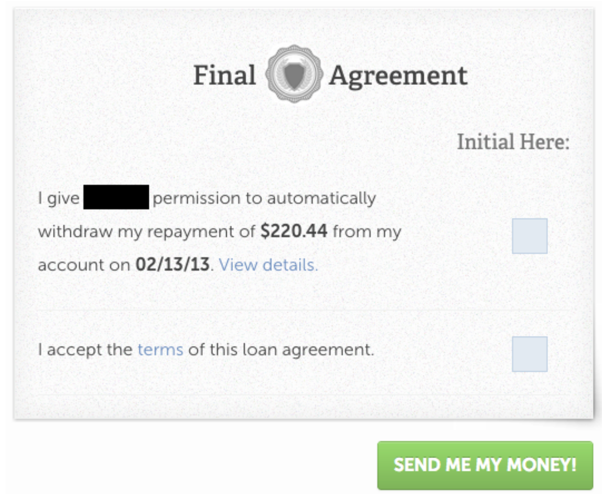


Panel (B.1.7): *Screen #7: Honor Pledge/Treatment Screen*

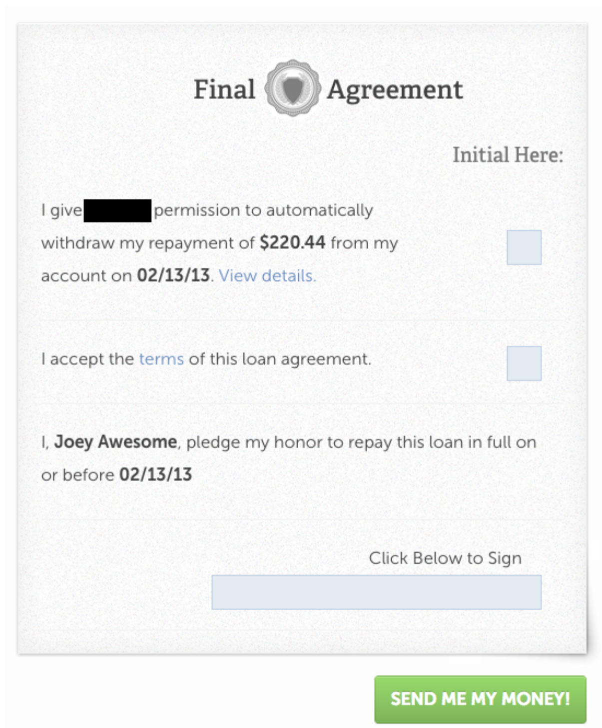


Panel (B.1.8): *Screen #8: Confirmation*

Appendix B.2: Control and Treatment Screens

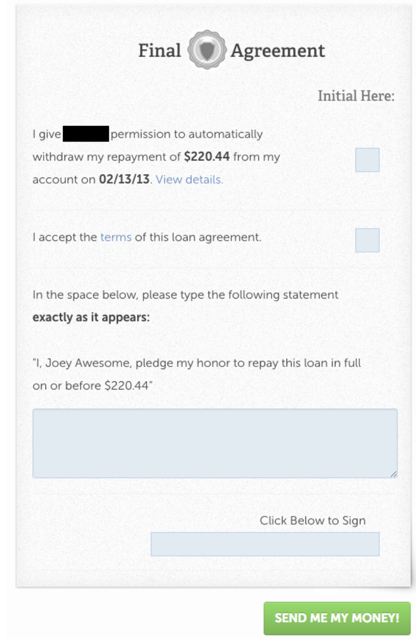


Panel (B.2.1): *Control Final Pre-Approval Screen:*

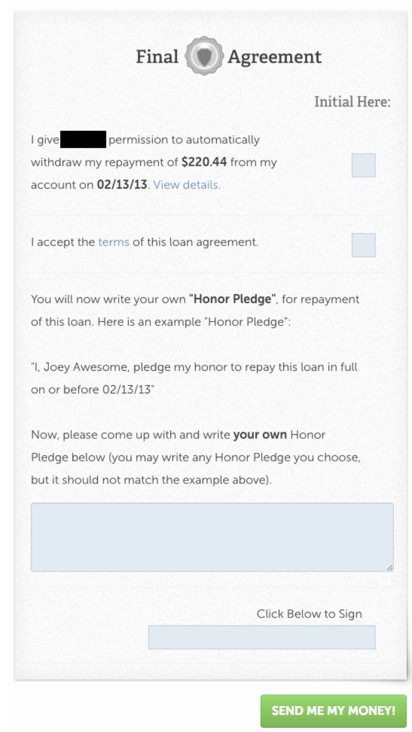


Panel (B.2.2): *"Simple" Honor Code Final Pre-Approval Screen:*

Appendix B.2: Control and Treatment Screens (continued)

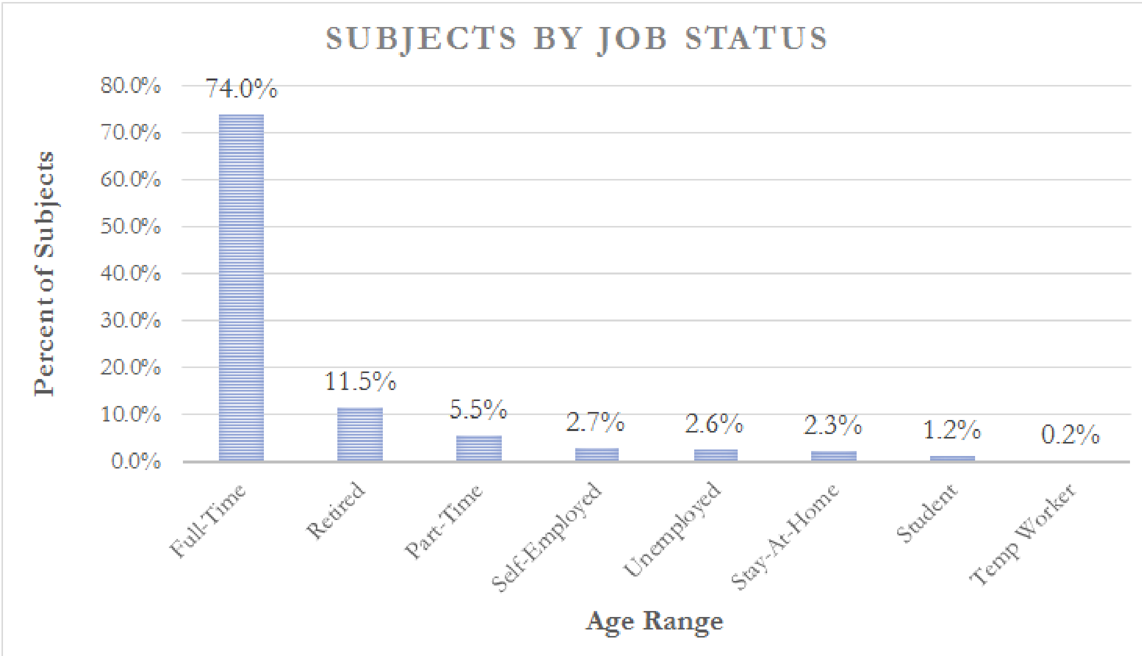


Panel (B.2.3): *"Copy" Honor Code Final Pre-Approval Screen:*

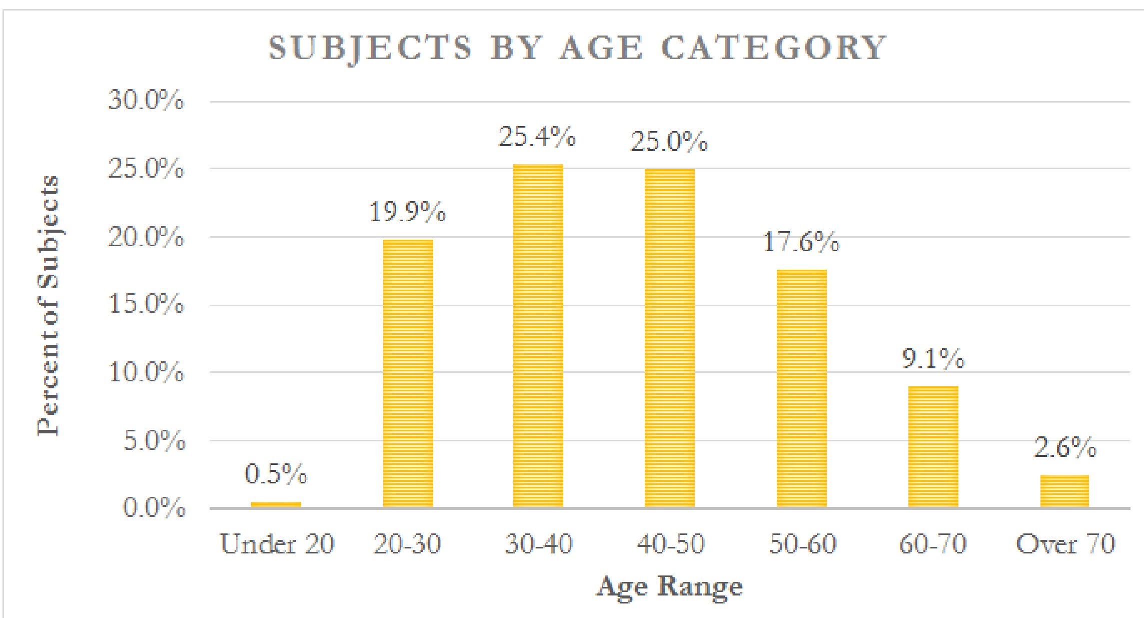


Panel (B.2.4): *"Write-In" Honor Code Final Pre-Approval Screen:*

Appendix B.3: Demographics Charts

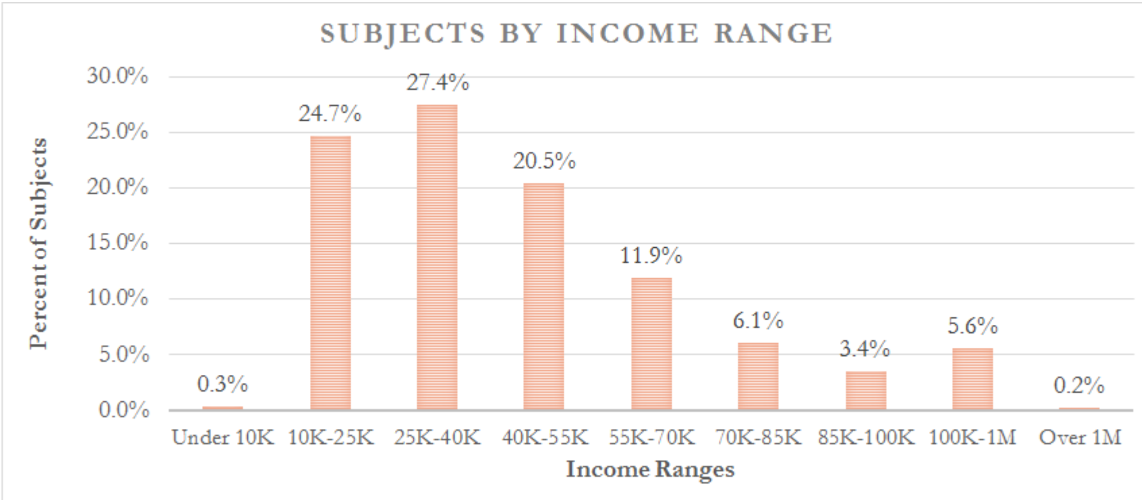


Panel (B.3.1): *Subjects by Job Status*



Panel (B.3.2): *Subjects by Age Category*

Appendix B.3: Demographics Charts (continued)



Panel (B.3.3): *Subjects by Income Range*

Table B.1: *Demographic Variables by Treatment Group*

	Control	Simple	Simple REM	Copy	Copy REM	Write-In	Write-In REM
Age	42.60	42.40	42.86	42.34	41.97	43.37	42.87
	(13.15)	(13.68)	(13.52)	(12.85)	(13.22)	(13.40)	(13.66)
Probability Male (see notes)	0.427	0.460	0.435	0.461	0.466	0.447	0.421
	(0.417)	(0.417)	(0.414)	(0.418)	(0.419)	(0.422)	(0.416)
Dummy for Male (see notes)	0.411	0.453	0.422	0.458	0.467	0.423	0.402
	(0.493)	(0.498)	(0.495)	(0.499)	(0.499)	(0.495)	(0.491)
Employed Full-Time	0.726	0.727	0.748	0.751	0.739	0.753	0.735
	(0.446)	(0.446)	(0.434)	(0.432)	(0.440)	(0.432)	(0.442)
Unemployed	0.0233	0.0296	0.0213	0.0291	0.0235	0.0302	0.0265
	(0.151)	(0.170)	(0.145)	(0.168)	(0.152)	(0.171)	(0.161)
Self-Reported Income	47116.7	48782.4	60914.8	67396.8	50158.2	50212.5	45891.3
	(42169.8)	(56177.2)	(245641.5)	(291373.8)	(154274.8)	(58139.9)	(93633.5)
Loan Amount at Inception	219.0	219.5	219.1	221.8	217.7	218.9	218.0
	(37.68)	(38.32)	(38.42)	(37.38)	(40.97)	(37.95)	(40.95)
Days of Loan at Inception	19.36	19.78	19.41	19.03	19.35	19.08	19.06
	(7.675)	(7.935)	(7.748)	(7.649)	(7.819)	(7.698)	(7.577)
N	729	675	656	688	723	695	717

Means, with standard deviations in parentheses, of key demographic variables, across treatment groups.
ProbMale reports the mean value for the variable that estimates likelihood of a given subject being male based on the first name of the subject.
Male uses the Probability Male variable. It reports the mean value of a dummy variable for male that takes a value of 1 for anyone over 99 percent likely to be male, and a value of 0 for anyone under 1 percent likely to be male. This therefore omits subjects who are 1-99 percent likely to be male.

Table B.2: *Randomization Checks*

	(1) Age	(2) ProbMale	(3) Male	(4) Emp. Full-Time	(5) Unemp.	(6) Income	(7) Loan Amount	(8) Loan Days
Simple	-0.197	0.0325	0.0421	0.00176	0.00631	1665.7	0.497	0.428
	(0.717)	(0.0228)	(0.0329)	(0.0238)	(0.00860)	(2667.3)	(2.030)	(0.417)
Simple+Reminder	0.257	0.00811	0.0113	0.0228	-0.00198	13798.1	0.0700	0.0502
	(0.718)	(0.0230)	(0.0334)	(0.0237)	(0.00795)	(9716.6)	(2.049)	(0.415)
Copy	-0.265	0.0338	0.0469	0.0258	0.00575	20280.1*	2.774	-0.323
	(0.691)	(0.0227)	(0.0326)	(0.0234)	(0.00851)	(11217.7)	(1.995)	(0.407)
Copy+Reminder	-0.630	0.0385*	0.0561*	0.0129	0.000194	3041.5	-1.295	-0.00258
	(0.692)	(0.0224)	(0.0323)	(0.0233)	(0.00794)	(5946.5)	(2.066)	(0.407)
WriteIn	0.773	0.0198	0.0115	0.0269	0.00690	3095.8	-0.107	-0.272
	(0.704)	(0.0227)	(0.0322)	(0.0233)	(0.00857)	(2702.4)	(2.005)	(0.408)
WriteIn+Reminder	0.266	-0.00584	-0.00943	0.00936	0.00318	-1225.3	-0.979	-0.294
	(0.705)	(0.0224)	(0.0320)	(0.0234)	(0.00820)	(3829.8)	(2.070)	(0.401)
F_value	0.815	1.207	1.198	0.451	0.344	1.102	0.822	0.799
P_value	0.558	0.299	0.304	0.845	0.914	0.358	0.553	0.571

Standard errors in parentheses.

This table presents the regressions of borrower characteristics on dummies for the six treatment groups.

F_value and P_value report the f-statistic and p-value associated with an f-test for joint significance of means for treatment coefficients.

ProbMale reports the mean value for the variable that estimates likelihood of a given subject being male based on the first name of the subject.

Male uses the Probability Male variable. It reports the mean value of a dummy variable for male that takes a value of 1 for anyone over 99% likely to be male, and a value of 0 for anyone under 1 percent likely to be male. This therefore omits individuals who are 1-99% likely to be male.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.3: *Average Treatment Effects on Outcome Variables: All Subjects*

	(1) Paid Off	(2) Perfect Payer	(3) Days to Pay	(4) Ever Overdue	(5) PayPlan	(6) PayPlan Paid
Simple	0.019	0.0056	0.74	-0.0036	0.019	0.11
	(0.022)	(0.026)	(0.93)	(0.023)	(0.052)	(0.067)
Simple-HP	0.0039	0.0067	0.35	0.012	0.023	0.056
	(0.022)	(0.026)	(0.92)	(0.023)	(0.051)	(0.062)
Copy	0.025	0.029	0.29	-0.014	-0.0085	0.047
	(0.021)	(0.026)	(0.85)	(0.023)	(0.053)	(0.066)
Copy-HP	0.012	0.019	-0.11	0.0018	-0.0086	0.036
	(0.021)	(0.026)	(0.82)	(0.023)	(0.052)	(0.066)
WriteIn	0.0078	0.020	-0.067	-0.00075	-0.015	0.057
	(0.022)	(0.026)	(0.79)	(0.023)	(0.051)	(0.067)
WriteIn-HP	0.0069	0.0038	0.71	0.0028	-0.058	0.059
	(0.021)	(0.026)	(0.98)	(0.023)	(0.051)	(0.068)
Constant	0.78***	0.29**	8.75***	0.11	1.24***	-1.29***
	(0.094)	(0.12)	(3.20)	(0.098)	(0.26)	(0.36)
Observations	4883	4883	3877	4883	1229	591
R^2	0.028	0.030	0.170	0.031	0.067	0.102
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Loan Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses.

All regression estimates above include controls for loan characteristics and borrower demographics.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.4: *Average Treatment Effects on Outcome Variables: All Subjects (Treatments Grouped)*

	(1) Paid Off	(2) Perfect Payer	(3) Days to Pay	(4) Ever Overdue	(5) PayPlan	(6) PayPlan Paid
All Simple	0.012	0.0062	0.55	0.0042	0.021	0.081
	(0.019)	(0.023)	(0.79)	(0.020)	(0.044)	(0.055)
All Copy	0.018	0.024	0.085	-0.0058	-0.0082	0.041
	(0.018)	(0.022)	(0.73)	(0.020)	(0.045)	(0.057)
All WriteIn	0.0073	0.012	0.33	0.0011	-0.037	0.058
	(0.019)	(0.022)	(0.76)	(0.020)	(0.044)	(0.057)
Constant	0.78***	0.28**	8.85***	0.12	1.24***	-1.29***
	(0.093)	(0.12)	(3.21)	(0.098)	(0.26)	(0.36)
Observations	4883	4883	3877	4883	1229	591
R^2	0.028	0.030	0.170	0.031	0.067	0.101
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Loan Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.5: *Average Treatment Effects on Outcome Variables: All Subjects (Treatments Grouped, Reminder Effect Isolated)*

	(1) Paid Off	(2) Perfect Payer	(3) Days to Pay	(4) Ever Overdue	(5) PayPlan	(6) PayPlan Paid
All Simple	0.016	0.010	0.55	-0.0015	0.028	0.091
	(0.020)	(0.024)	(0.83)	(0.021)	(0.047)	(0.059)
All Copy	0.023	0.029	0.084	-0.012	-0.0012	0.052
	(0.019)	(0.024)	(0.78)	(0.021)	(0.047)	(0.060)
All WriteIn	0.012	0.016	0.32	-0.0048	-0.030	0.068
	(0.020)	(0.024)	(0.76)	(0.021)	(0.047)	(0.060)
Treatment * Reminder	-0.0096	-0.0087	0.0028	0.012	-0.014	-0.020
	(0.012)	(0.015)	(0.52)	(0.013)	(0.030)	(0.039)
Constant	0.78***	0.29**	8.85***	0.11	1.24***	-1.29***
	(0.094)	(0.12)	(3.23)	(0.098)	(0.26)	(0.36)
Observations	4883	4883	3877	4883	1229	591
R^2	0.028	0.030	0.170	0.031	0.067	0.101
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Loan Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.6: *Average Treatment Effects on Outcome Variables: All Subjects, Richest Quintile*

	(1) Paid Off	(2) Perfect Payer	(3) Days to Pay	(4) Ever Overdue	(5) PayPlan	(6) PayPlan Paid
Simple	0.021	0.10*	-0.61	0.0066	0.022	0.11
	(0.049)	(0.057)	(3.85)	(0.052)	(0.11)	(0.14)
Simple-HP	-0.0021	0.026	-3.08	0.016	0.12	-0.17
	(0.049)	(0.059)	(3.29)	(0.053)	(0.11)	(0.13)
Copy	0.015	0.11*	-4.01	-0.036	-0.090	-0.092
	(0.050)	(0.060)	(2.91)	(0.052)	(0.12)	(0.16)
Copy-HP	-0.044	0.038	-4.42	0.049	-0.062	-0.19
	(0.050)	(0.058)	(2.98)	(0.052)	(0.11)	(0.13)
WriteIn	-0.028	0.013	-3.55	0.048	0.049	-0.11
	(0.051)	(0.059)	(2.89)	(0.054)	(0.11)	(0.13)
WriteIn-HP	0.029	0.075	-5.55**	-0.031	-0.039	-0.044
	(0.048)	(0.058)	(2.72)	(0.051)	(0.11)	(0.15)
Constant	1.24***	0.82***	6.81*	-0.34**	0.32	-1.57*
	(0.13)	(0.23)	(3.67)	(0.13)	(0.93)	(0.83)
Observations	992	992	760	992	280	151
R^2	0.052	0.057	0.113	0.059	0.130	0.147
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Loan Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses.

All regression estimates above include controls for loan characteristics and borrower demographics.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.7: *Average Treatment Effects on Outcome Variables: All Subjects, Second Richest Quintile*

	(1) Paid Off	(2) Perfect Payer	(3) Days to Pay	(4) Ever Overdue	(5) PayPlan	(6) PayPlan Paid
Simple	0.015	-0.044	0.95	0.024	-0.11	0.15
	(0.055)	(0.062)	(1.65)	(0.058)	(0.11)	(0.13)
Simple-HP	0.071	0.094	0.72	-0.027	-0.15	0.23*
	(0.051)	(0.061)	(1.65)	(0.055)	(0.11)	(0.14)
Copy	0.050	0.010	3.27	0.0047	-0.097	0.22
	(0.051)	(0.060)	(2.45)	(0.054)	(0.11)	(0.14)
Copy-HP	0.051	0.040	0.91	-0.034	-0.16	0.041
	(0.049)	(0.058)	(1.14)	(0.051)	(0.11)	(0.13)
WriteIn	0.0097	0.024	1.41	0.019	-0.19*	0.22
	(0.053)	(0.060)	(1.61)	(0.055)	(0.11)	(0.15)
WriteIn-HP	-0.0050	-0.042	2.30	0.028	-0.20*	0.023
	(0.050)	(0.059)	(1.81)	(0.052)	(0.11)	(0.14)
Constant	2.06	2.75*	-7.20	-1.40	-1.16	2.09
	(1.33)	(1.54)	(37.4)	(1.40)	(2.95)	(4.84)
Observations	978	978	746	978	282	138
R^2	0.050	0.046	0.240	0.054	0.122	0.178
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Loan Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses.

All regression estimates above include controls for loan characteristics and borrower demographics.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.8: *Average Treatment Effects on Outcome Variables: All Subjects, Middle Quintile*

	(1) Paid Off	(2) Perfect Payer	(3) Days to Pay	(4) Ever Overdue	(5) PayPlan	(6) PayPlan Paid
Simple	0.062	0.012	1.75	-0.049	0.11	-0.055
	(0.049)	(0.059)	(1.36)	(0.050)	(0.13)	(0.17)
Simple-HP	0.0037	-0.041	2.41	0.017	0.15	0.020
	(0.053)	(0.061)	(2.31)	(0.054)	(0.11)	(0.13)
Copy	0.077	0.0044	2.07	-0.052	0.17	-0.0099
	(0.047)	(0.058)	(1.49)	(0.049)	(0.13)	(0.16)
Copy-HP	0.020	0.0080	1.01	-0.015	0.057	-0.053
	(0.052)	(0.060)	(1.34)	(0.053)	(0.13)	(0.18)
WriteIn	0.12***	0.039	1.60	-0.11**	0.16	0.054
	(0.045)	(0.059)	(1.43)	(0.047)	(0.14)	(0.23)
WriteIn-HP	0.018	-0.023	5.98**	0.016	0.082	0.18
	(0.049)	(0.057)	(2.97)	(0.050)	(0.11)	(0.16)
Constant	-0.41	-1.06	53.4	1.76	-1.97	0.63
	(1.67)	(2.00)	(48.9)	(1.74)	(4.38)	(5.67)
Observations	961	961	761	961	226	104
R^2	0.047	0.043	0.202	0.054	0.108	0.175
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Loan Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses.

All regression estimates above include controls for loan characteristics and borrower demographics.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.9: *Average Treatment Effects on Outcome Variables: All Subjects, Second Poorest Quintile*

	(1) Paid Off	(2) Perfect Payer	(3) Days to Pay	(4) Ever Overdue	(5) PayPlan	(6) PayPlan Paid
Simple	-0.023	-0.015	0.94	0.044	0.12	0.12
	(0.051)	(0.060)	(1.59)	(0.055)	(0.12)	(0.17)
Simple-HP	0.0042	0.0093	0.19	0.0049	0.14	0.086
	(0.051)	(0.061)	(1.72)	(0.055)	(0.12)	(0.16)
Copy	0.0066	0.015	0.51	-0.0099	0.027	-0.056
	(0.049)	(0.058)	(1.37)	(0.052)	(0.12)	(0.17)
Copy-HP	0.058	0.075	0.65	-0.044	0.11	0.20
	(0.046)	(0.057)	(2.20)	(0.050)	(0.12)	(0.19)
WriteIn	-0.0023	0.018	-0.63	-0.0056	-0.031	0.042
	(0.050)	(0.059)	(1.41)	(0.052)	(0.12)	(0.16)
WriteIn-HP	0.053	0.043	1.20	-0.044	0.048	0.22
	(0.048)	(0.059)	(1.96)	(0.051)	(0.13)	(0.22)
Constant	1.11	0.62	34.2	0.25	-0.43	-5.35
	(1.34)	(1.66)	(63.7)	(1.44)	(3.10)	(5.65)
Observations	989	989	784	989	250	108
R^2	0.036	0.046	0.188	0.045	0.183	0.277
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Loan Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses.

All regression estimates above include controls for loan characteristics and borrower demographics.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.10: *Average Treatment Effects on Outcome Variables: All Subjects, Poorest Quintile*

	(1) Paid Off	(2) Perfect Payer	(3) Days to Pay	(4) Ever Overdue	(5) PayPlan	(6) PayPlan Paid
Simple	0.019	-0.024	-0.072	-0.045	-0.074	-0.0031
	(0.037)	(0.057)	(0.94)	(0.044)	(0.16)	(0.24)
Simple-HP	-0.051	-0.022	1.34	0.042	-0.17	0.22
	(0.041)	(0.057)	(1.38)	(0.048)	(0.14)	(0.23)
Copy	-0.017	0.034	-0.51	0.010	-0.10	-0.067
	(0.039)	(0.056)	(1.08)	(0.046)	(0.16)	(0.21)
Copy-HP	-0.023	-0.041	1.03	0.041	-0.11	0.15
	(0.040)	(0.058)	(1.27)	(0.049)	(0.14)	(0.21)
WriteIn	-0.049	0.025	0.57	0.031	-0.15	0.044
	(0.040)	(0.055)	(1.60)	(0.047)	(0.14)	(0.22)
WriteIn-HP	-0.071	-0.038	-0.51	0.059	-0.094	-0.17
	(0.047)	(0.062)	(1.06)	(0.052)	(0.15)	(0.21)
Constant	0.81***	0.52	15.0*	0.13	1.59**	-2.13
	(0.30)	(0.40)	(8.15)	(0.33)	(0.80)	(1.82)
Observations	963	963	826	963	191	90
R^2	0.029	0.045	0.332	0.036	0.099	0.199
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Loan Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses.

All regression estimates above include controls for loan characteristics and borrower demographics.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.11: *Average Treatment Effects: Age Interactions*

	(1) Paid Off	(2) Perf. Payer	(3) Days to Pay	(4) Ever Overdue	(5) PayPlan	(6) PP Paid
Simple*Age	-0.0017	-0.0032*	-0.065	0.00048	-0.0064	-0.00082
	(0.0016)	(0.0019)	(0.080)	(0.0017)	(0.0039)	(0.0054)
Simple-HP*Age	0.00028	-0.0016	0.037	0.00026	-0.0028	0.0066
	(0.0017)	(0.0020)	(0.082)	(0.0018)	(0.0038)	(0.0049)
Copy*Age	0.00059	-0.00052	-0.049	-0.00087	-0.0021	0.0018
	(0.0016)	(0.0020)	(0.056)	(0.0017)	(0.0044)	(0.0053)
Copy-HP*Age	0.00032	-0.0015	0.0034	-0.00035	-0.0058	0.0022
	(0.0017)	(0.0019)	(0.079)	(0.0017)	(0.0039)	(0.0050)
WriteIn*Age	-0.0010	-0.0038*	-0.025	0.00037	0.0021	-0.0017
	(0.0017)	(0.0020)	(0.065)	(0.0018)	(0.0038)	(0.0046)
WriteIn-HP*Age	-0.00029	-0.0023	-0.044	0.00057	-0.00094	0.0058
	(0.0016)	(0.0019)	(0.057)	(0.0017)	(0.0039)	(0.0053)
Simple	0.093	0.14	3.49	-0.024	0.29*	0.13
	(0.073)	(0.087)	(3.79)	(0.077)	(0.17)	(0.23)
Simple-HP	-0.0082	0.074	-1.25	0.00100	0.14	-0.22
	(0.077)	(0.089)	(3.50)	(0.080)	(0.17)	(0.21)
Copy	0.00011	0.052	2.41	0.023	0.080	-0.031
	(0.074)	(0.088)	(2.31)	(0.077)	(0.18)	(0.23)
Copy-HP	-0.0010	0.083	-0.24	0.016	0.23	-0.059
	(0.075)	(0.087)	(3.18)	(0.078)	(0.17)	(0.21)
WriteIn	0.052	0.18**	0.99	-0.017	-0.11	0.13
	(0.076)	(0.088)	(2.81)	(0.080)	(0.17)	(0.21)
WriteIn-HP	0.019	0.10	2.62	-0.022	-0.021	-0.19
	(0.072)	(0.086)	(2.76)	(0.076)	(0.17)	(0.22)
Age	0.0018	0.0017	0.049	-0.0017	0.0025	-0.0010
	(0.0012)	(0.0014)	(0.043)	(0.0013)	(0.0029)	(0.0034)
Constant	0.77***	0.22	7.88**	0.12	1.15***	-1.18***
	(0.10)	(0.13)	(3.21)	(0.11)	(0.28)	(0.37)
Observations	4883	4883	3877	4883	1229	591
R^2	0.029	0.031	0.171	0.031	0.074	0.110
Dem. Controls	Yes	Yes	Yes	Yes	Yes	Yes
Loan Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.12: *Average Treatment Effects: Gender Interactions*

	(1) Paid Off	(2) Perf. Payer	(3) Days to Pay	(4) Ever Overdue	(5) PayPlan	(6) PP Paid
Male	-0.027	0.045	-1.27	0.0084	-0.049	-0.049
	(0.037)	(0.045)	(0.90)	(0.038)	(0.099)	(0.11)
Simple*Male	-0.016	-0.056	2.61	0.048	-0.18	0.16
	(0.051)	(0.065)	(2.26)	(0.054)	(0.14)	(0.18)
Simple-HP*Male	0.023	-0.027	4.74**	0.018	0.17	0.22
	(0.054)	(0.067)	(1.97)	(0.058)	(0.14)	(0.16)
Copy*Male	0.028	-0.044	1.33	0.0089	-0.12	0.23
	(0.052)	(0.064)	(1.50)	(0.055)	(0.14)	(0.18)
Copy-HP*Male	0.0033	-0.070	1.07	0.017	-0.024	0.076
	(0.053)	(0.064)	(1.77)	(0.056)	(0.14)	(0.16)
WriteIn*Male	0.068	-0.0041	1.40	-0.038	-0.077	0.21
	(0.051)	(0.064)	(1.69)	(0.054)	(0.15)	(0.17)
WriteIn-HP*Male	0.044	0.049	-0.97	-0.050	-0.014	-0.10
	(0.053)	(0.065)	(1.55)	(0.056)	(0.14)	(0.16)
Simple	0.040	0.045	0.34	-0.046	0.12	0.059
	(0.032)	(0.043)	(0.97)	(0.035)	(0.096)	(0.12)
Simple-HP	-0.0046	0.037	-1.84**	0.011	-0.081	-0.049
	(0.034)	(0.043)	(0.83)	(0.037)	(0.091)	(0.10)
Copy	0.0040	0.034	0.12	0.00030	-0.041	0.028
	(0.034)	(0.043)	(1.02)	(0.036)	(0.092)	(0.12)
Copy-HP	-0.018	0.028	0.018	0.023	-0.045	0.022
	(0.034)	(0.042)	(1.42)	(0.037)	(0.092)	(0.11)
WriteIn	-0.020	0.0056	0.28	0.019	0.0023	0.017
	(0.033)	(0.042)	(1.10)	(0.035)	(0.091)	(0.11)
WriteIn-HP	-0.034	-0.053	1.04	0.050	-0.100	0.13
	(0.033)	(0.042)	(1.27)	(0.036)	(0.084)	(0.11)
Constant	0.79***	0.29*	10.5**	0.12	1.15***	-0.93**
	(0.12)	(0.16)	(4.51)	(0.13)	(0.34)	(0.47)
Observations	3133	3133	2545	3133	719	372
R^2	0.025	0.032	0.186	0.026	0.084	0.124
Dem. Controls	Yes	Yes	Yes	Yes	Yes	Yes
Loan Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses.

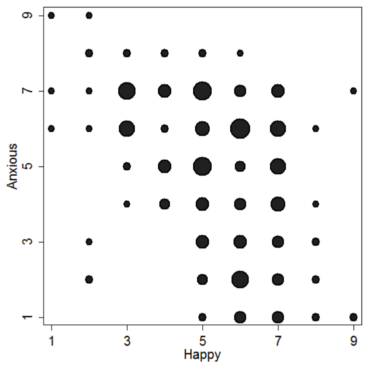
This regression omits individuals who the gender first-name algorithm could not identify as male or female with greater than 99 percent probability.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

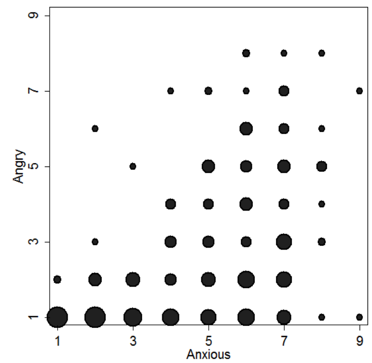
Appendix C

Appendix to Chapter 3

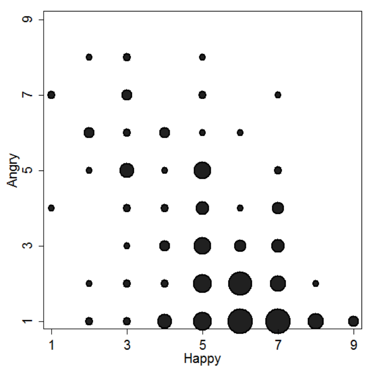
Appendix C.1: Test Score Study: Correlation between Current Emotions



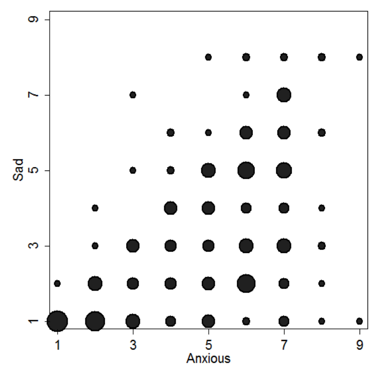
Panel (C.1.1): *Anxious vs. Happy*



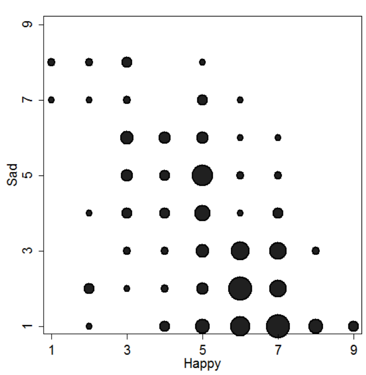
Panel (C.1.2): *Angry vs. Anxious*



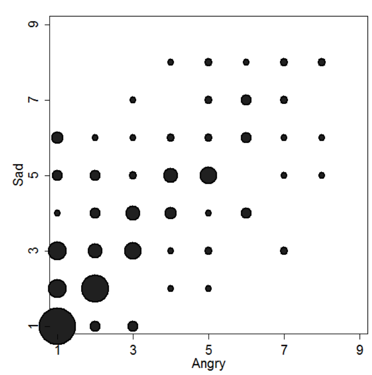
Panel (C.1.3): *Angry vs. Happy*



Panel (C.1.4): *Sad vs. Anxious*

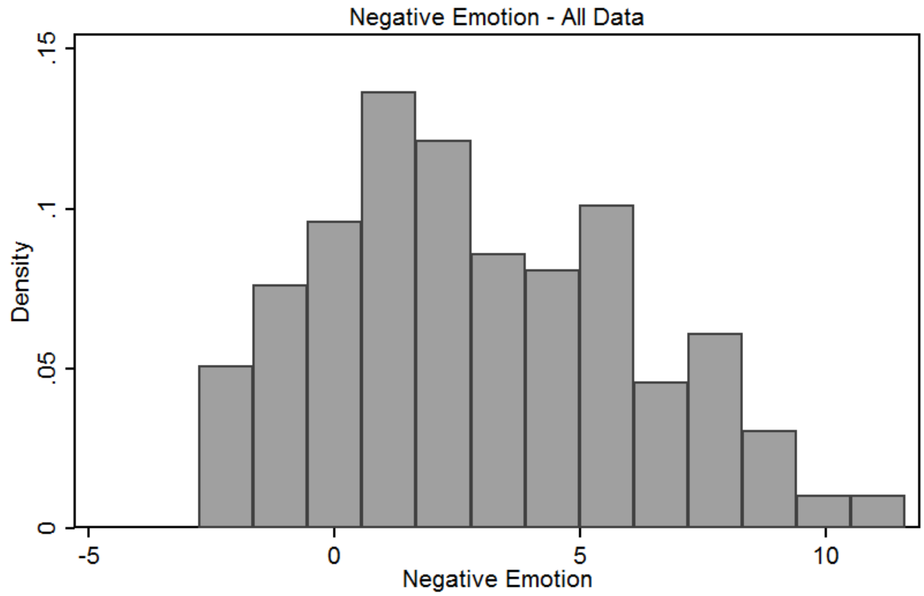


Panel (C.1.5): *Sad vs. Happy*

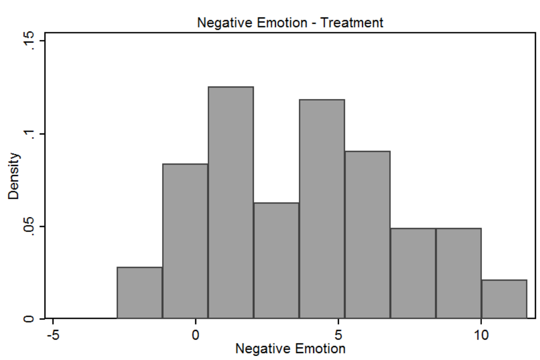


Panel (C.1.6): *Sad vs. Angry*

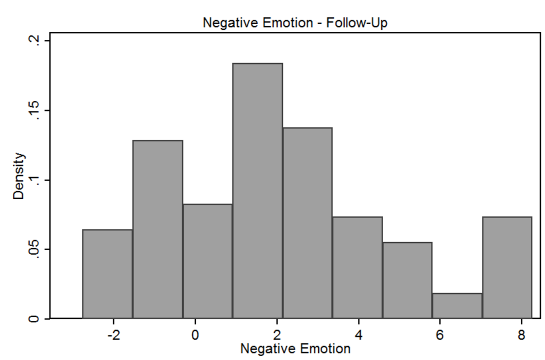
Appendix C.2: Test Score Study: Negative Emotion Variable Distribution



Panel (C.2.1): *All Data*

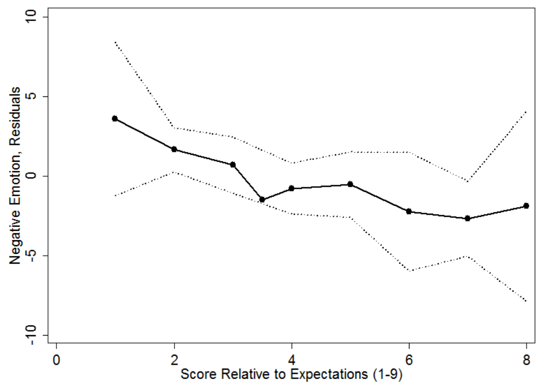


Panel (C.2.2): *Treatment Only*

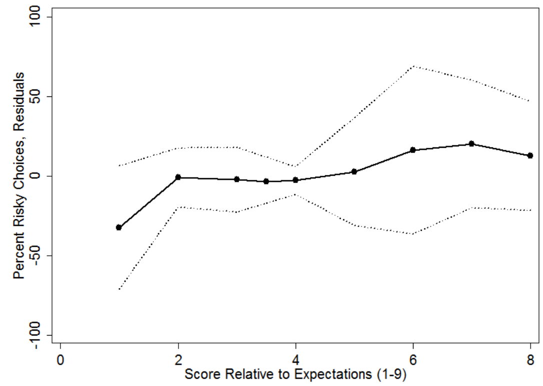


Panel (C.2.3): *Follow-Up Only*

Appendix C.3: Test Score Study: Outcomes by Exam Performance



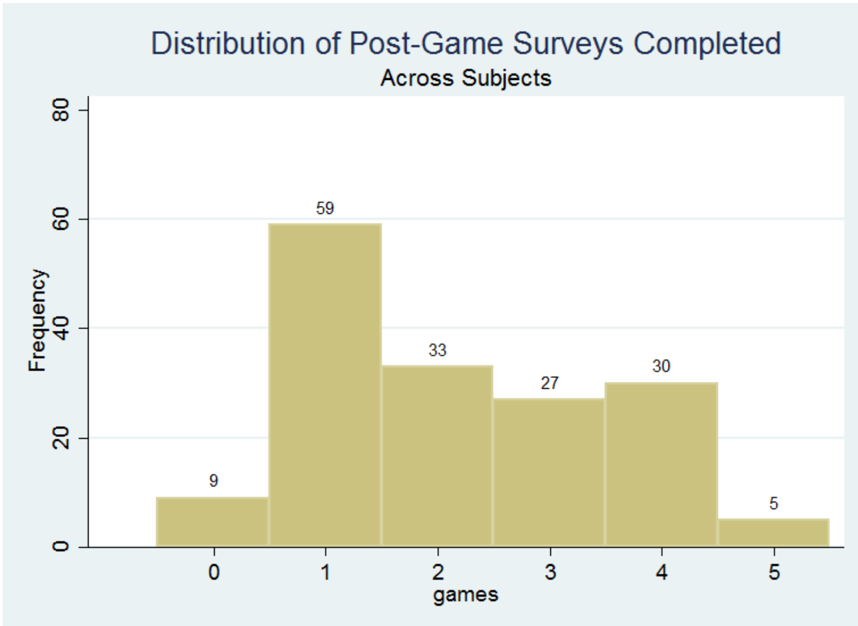
Panel (C.3.1): *Negative Emotion*



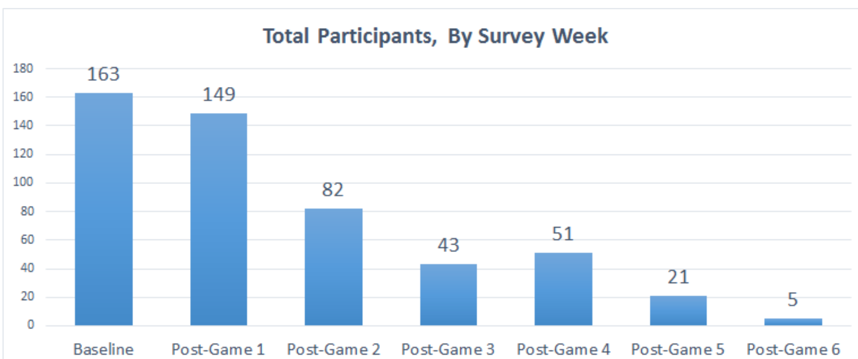
Panel (C.3.2): *Percent Risky Choices*

Note: The x-axis in each panel represents performance on the midterm relative to expectations. The y-axis in each panel represents residuals from a first-differenced regression of the specified outcome variable on absolute midterm score and a time period variable. Each panel plots means and 95 percent confidence intervals for the residuals for each value of the relative to expectations variable.

Appendix C.4: NFL Fans Study: Descriptive Bar Charts

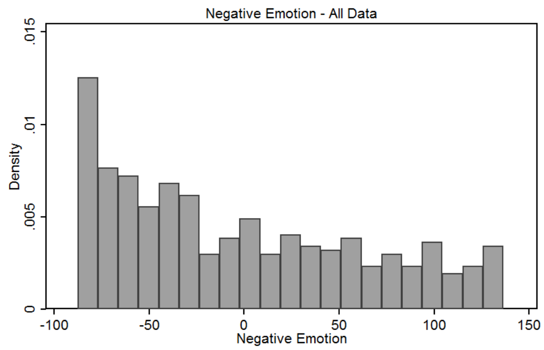


Panel (C.4.1): *Distribution of Post-Game Surveys Completed*

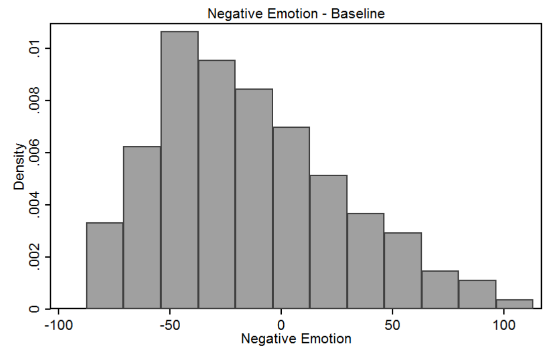


Panel (C.4.2): *Participants by Survey Week*

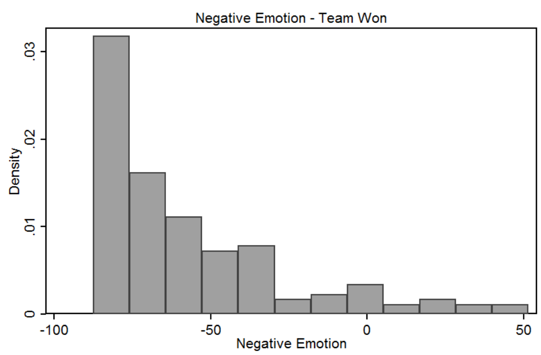
Appendix C.5: NFL Fan Study: Negative Emotion Variable Distribution



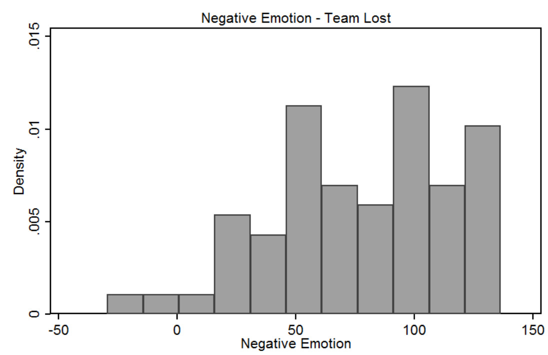
Panel (C.5.1): *All Data*



Panel (C.5.2): *Baseline Only*

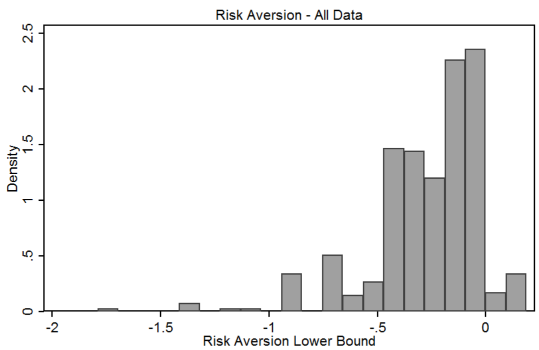


Panel (C.5.3): *Team Won Only*

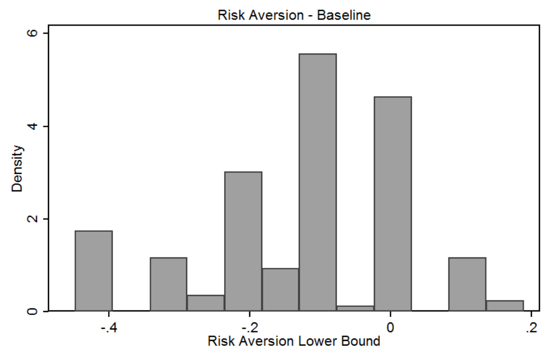


Panel (C.5.4): *Team Lost Only*

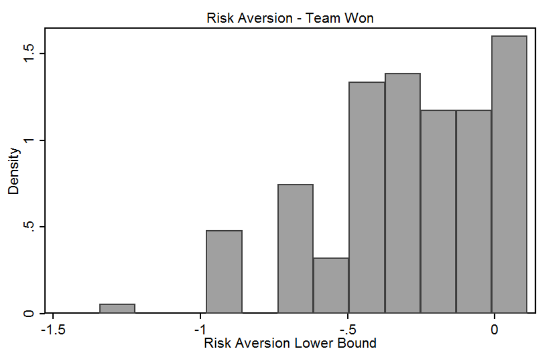
Appendix C.6: NFL Fan Study: Risk Aversion Variable Distribution



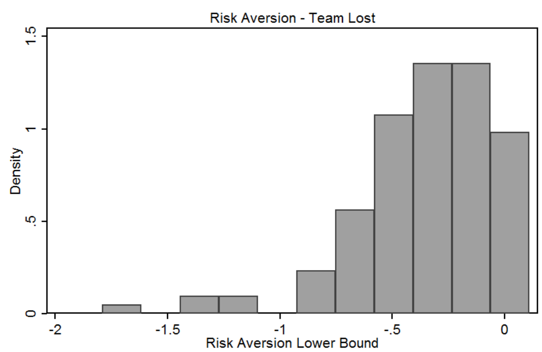
Panel (C.6.1): *All Data*



Panel (C.6.2): *Baseline Only*

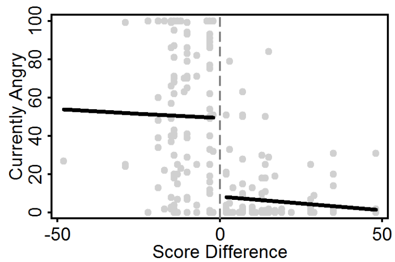


Panel (C.6.3): *Team Won Only*

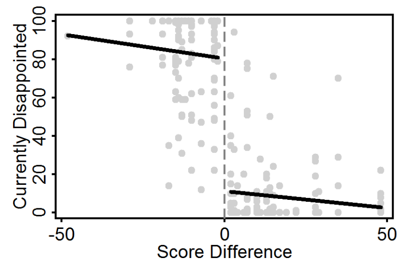


Panel (C.6.4): *Team Lost Only*

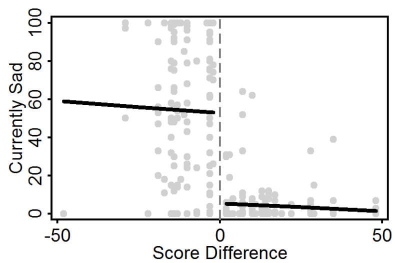
Appendix C.7: NFL Fan Study: Emotion vs. Score Gap (Own Team - Opponent)



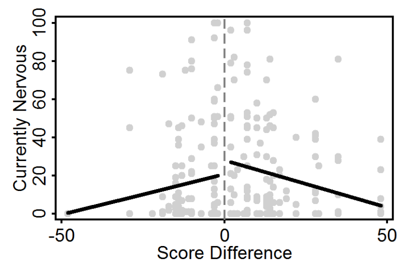
Panel (C.7.1): *Currently Angry*



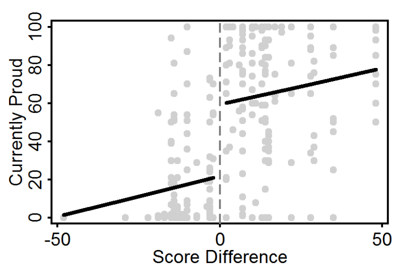
Panel (C.7.2): *Currently Disappointed*



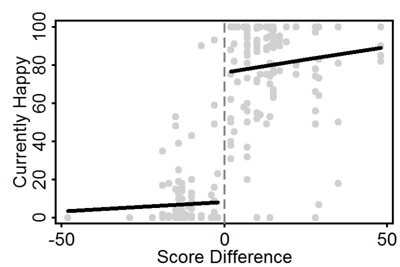
Panel (C.7.3): *Currently Sad*



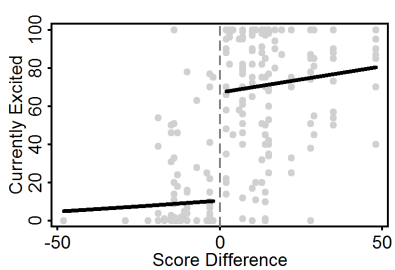
Panel (C.7.4): *Currently Nervous*



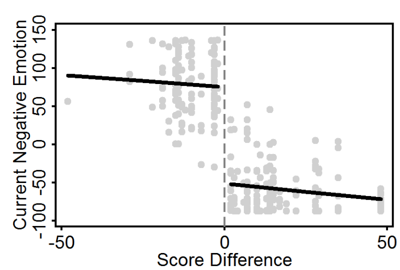
Panel (C.7.5): *Currently Proud*



Panel (C.7.6): *Currently Happy*

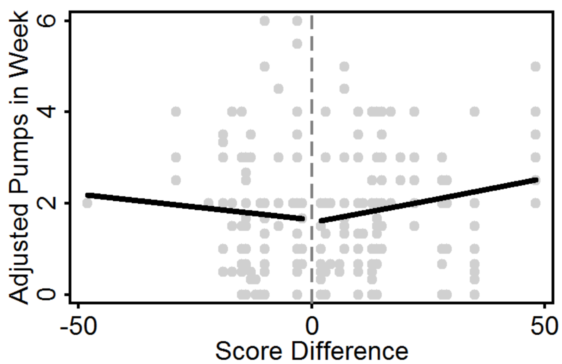


Panel (C.7.7): *Currently Excited*

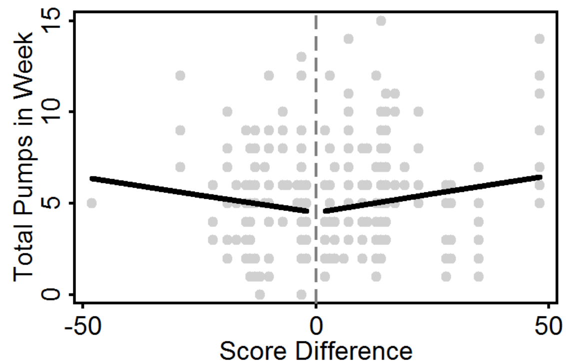


Panel (C.7.8): *Current Negative Emotion*

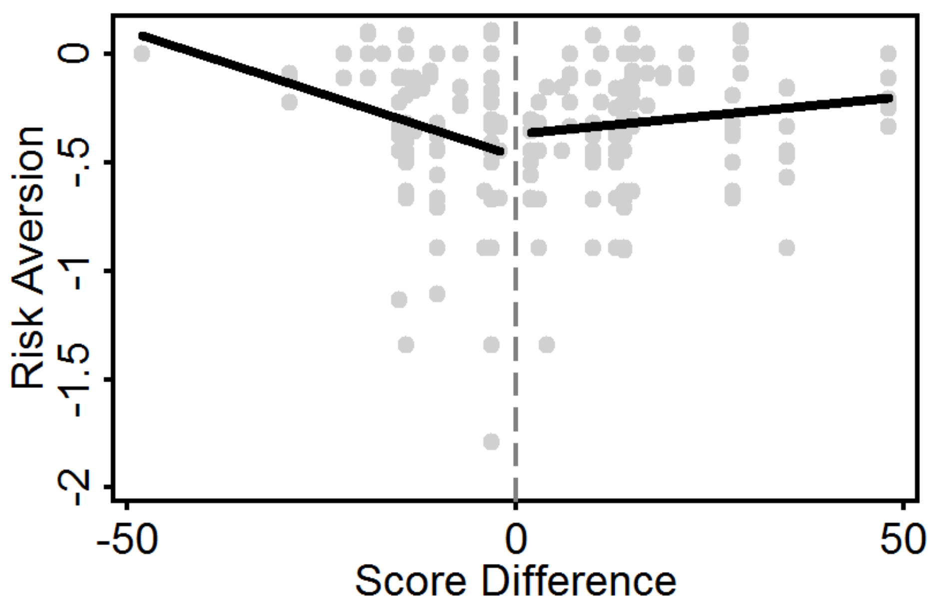
Appendix C.8: NFL Fan Study: Risk vs. Score Gap (Own Team - Opponent)



Panel (C.8.1): *Adjusted Pumps*



Panel (C.8.2): *Total Pumps per Week*



Panel (C.8.3): *Risk Aversion*

Table C.1: *Test Score Study: Demographic Variables*

	All Participants
Female	0.678
	(0.470)
College Year (1-4)	2.689
	(0.697)
GPA	3.481
	(0.328)
Usually Happy (1-7)	5.278
	(1.272)
Usually Sad (1-7)	2.944
	(1.617)
Usually Anxious (1-7)	3.900
	(1.793)
Usually Angry (1-7)	2.078
	(1.211)
School is Important (1-7)	6.228
	(1.031)
Midterm Score (0-100)	74.68
	(15.49)
Perf. Relative to Expect. (1-9)	3.728
	(1.675)
Risk Percentage in Treatment	43.94
	(33.73)
Risk Percentage in Follow-Up	41.48
	(32.50)
N	90

This table presents demographic and outcome variables for student participants. Means are presented, with standard deviation in parentheses.

Table C.2: *Test Score Study: Principal Components of Emotion*

	Component 1	Component 2	Component 3	Component 4
Eigenvalues	2.629753	.5853373	.5099897	.2749197
Eigenvectors				
Happy	-.4744803	.4435037	.7546604	.0930628
Anxious	.4545746	.8584207	-.2035512	-.1226489
Angry	.5208673	-.2353913	.5418137	-.6162193
Sad	.5449107	-.1049251	.3090188	.7723797

This table presents the results of a principal components analysis of the relationship between current happiness, sadness, anxiety, and anger.
This table is based on both treatment and follow-up vectors of emotions for each participant. The total sample size is 179 observations.

Table C.3: *Test Score Study: Exam Performance and Participant Characteristics*

	(1) Female	(2) Years	(3) GPA	(4) O. Happy	(5) O. Sad	(6) O. Anxious	(7) O. Angry	(8) School Imp.	(9) Exam Imp.
Midterm Score (0-100)	0.00072	-0.0019	0.0086***	0.0026	0.0033	0.0034	0.0012	0.0062	0.021*
	(0.0036)	(0.0063)	(0.0026)	(0.0078)	(0.015)	(0.016)	(0.0097)	(0.0079)	(0.012)
Score Rel. to Exp. (1-9)	0.017	-0.020	0.0056	-0.14	0.077	0.24*	-0.031	-0.15	-0.18*
	(0.031)	(0.048)	(0.019)	(0.11)	(0.12)	(0.13)	(0.095)	(0.11)	(0.11)
Observations	90	90	90	90	90	90	90	90	90
R^2	0.006	0.006	0.176	0.031	0.010	0.057	0.002	0.045	0.052

Standard errors in parentheses.
This table shows the results of regressions of participant characteristics on midterm test performance (both absolute score and relative to expectations). The model variables are gender, years in school, GPA, four measures of how often subjects felt listed emotions, and two measures of how important participants deemed school and the midterm exam.
This table is intended as a balance check to assess if exam performance is independent of participant characteristics.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.4: *Test Score Study: The Impact of Test Scores on Specific Emotions*

	(1) Happy	(2) Happy	(3) Anxious	(4) Anxious	(5) Angry	(6) Angry	(7) Sad	(8) Sad
Midterm Score (0-100)	0.030***	0.029**	0.0046	0.0082	-0.024	-0.032	-0.019	-0.024
	(0.011)	(0.014)	(0.015)	(0.025)	(0.016)	(0.024)	(0.016)	(0.027)
Score Relative to Expectations (1-9)	0.42***	0.69***	-0.37***	-0.48**	-0.51***	-0.49***	-0.47***	-0.57***
	(0.11)	(0.11)	(0.12)	(0.21)	(0.13)	(0.17)	(0.12)	(0.19)
Follow-up Questionnaire		5.80***		-2.22		-4.98***		-4.65**
		(0.94)		(1.70)		(1.59)		(1.81)
Usually Happy (1-7)	0.32**							
	(0.12)							
Usually Anxious (1-7)			0.42***					
			(0.11)					
Usually Angry (1-7)					0.44**			
					(0.19)			
Usually Sad (1-7)							0.51***	
							(0.12)	
Observations	90	178	90	178	90	178	90	178
R^2	0.427	0.773	0.324	0.563	0.409	0.646	0.472	0.636
Demographics	Yes	No	Yes	No	Yes	No	Yes	No
Participant Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Includes Follow-up Data	No	Yes	No	Yes	No	Yes	No	Yes

Standard errors in parentheses.

This table shows the results of regressions of current emotions on midterm performance (both absolute score and relative to expectations. Each emotion variable has two columns above - one reporting results from the first survey with demographic controls (gender, years in school, and GPA), and one with data from both surveys with a dummy for follow-up, which is effectively a time fixed effect.

Note that the four dependent variables in models 1-8 are self-reported current emotions on a 0-100 scale.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.5: *Test Score Study: First-Stage IV Regression (Y=Negative Emotion)*

	(1) Neg. Emotion	(2) Neg. Emotion	(3) Neg. Emotion	(4) Neg. Emotion	(5) Neg. Emotion
Midterm Score (0-100)	-0.050**	-0.036*	-0.045**		-0.039
	(0.021)	(0.021)	(0.019)		(0.040)
Score Relative to Expectations (1-9)	-1.00***	-0.94***	-0.96***	-1.28***	-1.11***
	(0.17)	(0.17)	(0.16)	(0.23)	(0.27)
Follow-up Questionnaire			-8.83***	-6.58***	-8.89***
			(1.30)	(0.97)	(2.59)
Usually Happy (1-7)	-0.37*	-0.42**	-0.30**		
	(0.22)	(0.20)	(0.15)		
Usually Anxious (1-7)	0.39**	0.26	0.43***		
	(0.17)	(0.17)	(0.13)		
Usually Angry (1-7)	0.34*	0.30	0.30**		
	(0.19)	(0.21)	(0.14)		
Usually Sad (1-7)	0.32*	0.28	0.28**		
	(0.19)	(0.19)	(0.14)		
Observations	90	90	178	178	178
R^2	0.522	0.576	0.536	0.682	0.690
Demographics	No	Yes	Yes	No	No
Participant Fixed Effects	No	No	No	Yes	Yes
Includes Follow-up Data	No	No	Yes	Yes	Yes

Standard errors in parentheses.

This table shows the first-stage IV regression of negative emotion on exam performance. The negative emotion variable ranges from -2.75 to 11.62 in the data.

Models with and without demographic controls (gender, years in school, and GPA), fixed effects, and follow-up data are presented.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.6: *Test Score Study: Effect of Negative Emotions on Riskiness Percentage (Naive and IV Second-Stage)*

	(1) Naive	(2) Naive	(3) Naive	(4) Naive w/ FEs	(5) IV	(6) IV	(7) IV	(8) IV w/ FEs	(9) IV w/ FEs
Negative Emotion	-1.58	-0.71	-1.58	-1.12	-2.66**	-1.09	-2.83**	-4.34**	-3.82*
	(1.10)	(0.87)	(1.00)	(0.95)	(1.30)	(1.02)	(1.28)	(2.18)	(2.03)
Follow-up Questionnaire		-4.18	-5.25	-4.73		-4.85	-7.68	-10.5*	-9.60*
		(5.24)	(5.41)	(4.35)		(5.33)	(5.68)	(5.87)	(5.74)
Usually Happy (1-7)			1.82				1.67		
			(2.25)				(2.30)		
Usually Anxious (1-7)			3.66**				3.76**		
			(1.80)				(1.80)		
Usually Angry (1-7)			-3.73				-3.32		
			(2.50)				(2.47)		
Usually Sad (1-7)			0.76				1.30		
			(2.31)				(2.33)		
Observations	88	176	176	176	88	176	176	176	176
R^2	0.025	0.006	0.072	0.658	0.047	0.008	0.084	0.672	0.669
Demographics	No	No	Yes	No	No	No	Yes	No	No
Participant Fixed Effects	No	No	No	Yes	No	No	No	Yes	Yes
Includes Follow-up Data	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Instrument: Rel. to Expect.	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Instrument: Midterm Score	No	No	No	No	No	No	No	No	Yes

Standard errors in parentheses.

This table shows the results of regressions of the percentage of time a risk was taken on negative emotion, both in 'naive' form and with IV methods. Specifications 1-4 present 'naive' results, without using exam performance as an instrument for emotion, both without and with demographic controls (gender, years in school, and GPA), follow-up data, and fixed effects. Specifications 5-9 present results using exam performance as an instrument.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.7: *NFL Fans Study: Demographic Variables (Overall and By Team Quality)*

	All Fans	Playoff Fans	Chance Fans	Out Fans
Female	0.0307	0.0364	0.0282	0.0270
	(0.173)	(0.189)	(0.167)	(0.164)
Age 30+	0.644	0.582	0.690	0.649
	(0.480)	(0.498)	(0.466)	(0.484)
HH Income	72638.0	71363.6	76197.2	67702.7
	(27093.3)	(27053.7)	(24733.8)	(31037.5)
Educ: HS	0.0184	0.0182	0.0282	0
	(0.135)	(0.135)	(0.167)	(0)
Educ: Some College	0.196	0.218	0.197	0.162
	(0.398)	(0.417)	(0.401)	(0.374)
Educ: College Grad	0.558	0.509	0.521	0.703
	(0.498)	(0.505)	(0.503)	(0.463)
Educ: Grad Degree	0.227	0.255	0.254	0.135
	(0.420)	(0.440)	(0.438)	(0.347)
Fan Level (1-4)	3.521	3.509	3.577	3.432
	(0.661)	(0.690)	(0.601)	(0.728)
Team Watched: TV	14.45	14.33	14.42	14.70
	(2.857)	(2.951)	(2.984)	(2.504)
Team Watched: Live	2.123	2	1.986	2.568
	(2.526)	(2.681)	(2.207)	(2.863)
All Watched: Live	2.834	2.582	2.408	4.027
	(3.017)	(2.973)	(2.499)	(3.693)
Often Sad (0-100)	55.83	56.36	57.39	52.03
	(29.21)	(32.35)	(29.09)	(24.56)
Often Angry (0-100)	56.44	56.82	57.04	54.73
	(26.58)	(27.41)	(23.97)	(30.53)
Often Disapp. (0-100)	78.53	76.36	79.23	80.41
	(21.31)	(24.26)	(20.70)	(17.81)
Often Frust. (0-100)	71.47	65.45	72.54	78.38
	(24.67)	(29.46)	(22.02)	(19.69)
N	163	55	71	37

This table presents demographic variables for participants, both overall and by team quality. Playoff Fans are those rooting for teams that were guaranteed playoff spots, chance fans are those rooting for teams with a chance to make the playoffs, and out fans are those rooting for teams with no chance of making the playoffs.

Table C.8: *NFL Fans Study: Demographics and Team Quality (1 of 2)*

	(1) Female	(2) Age 30+	(3) Income	(4) HS grad	(5) Some College	(6) College Grad	(7) Some Grad School
Playoff Team	0.0093	-0.067	3660.9	0.018	0.056	-0.19*	0.12
	(0.037)	(0.10)	(5750.3)	(0.029)	(0.085)	(0.11)	(0.089)
Chance Team	0.0011	0.041	8494.5	0.028	0.035	-0.18*	0.12
	(0.035)	(0.098)	(5483.6)	(0.027)	(0.081)	(0.10)	(0.085)
Constant	0.027	0.65***	67702.7***	0	0.16**	0.70***	0.14*
	(0.029)	(0.079)	(4446.1)	(0.022)	(0.066)	(0.081)	(0.069)
N	163	163	163	163	163	163	163
R^2	0.001	0.010	0.016	0.007	0.003	0.025	0.014

Standard errors in parentheses.

Omitted group is fans with team eliminated from playoffs already entering study. Playoff Team is a dummy for if favorite team is in the playoffs entering the study. Chance Team is a dummy for if favorite team has a chance to make the playoffs entering the study. TeamTV reports the number of team games the fan watched on TV in previous season. TeamLive reports the number of team games the fan watched live in previous season. AllLive reports the number of NFL games the fan watched live in previous season.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.9: *NFL Fans Study: Demographics and Team Quality (2 of 2)*

	(1) Fan Level	(2) TeamTV	(3) TeamLive	(4) AllLive	(5) Often Sad	(6) Often Angry	(7) Often Disapp.	(8) Often Frus.
Playoff Team	0.077	-0.38	-0.57	-1.45**	4.34	2.09	-4.04	-12.9**
	(0.14)	(0.61)	(0.54)	(0.63)	(6.23)	(5.68)	(4.55)	(5.17)
Chance Team	0.15	-0.28	-0.58	-1.62***	5.37	2.31	-1.18	-5.84
	(0.13)	(0.58)	(0.51)	(0.60)	(5.94)	(5.42)	(4.34)	(4.93)
Constant	3.43***	14.7***	2.57***	4.03***	52.0***	54.7***	80.4***	78.4***
	(0.11)	(0.47)	(0.42)	(0.49)	(4.82)	(4.39)	(3.52)	(4.00)
N	163	163	163	163	163	163	163	163
R^2	0.007	0.002	0.009	0.047	0.005	0.001	0.006	0.039

Standard errors in parentheses.

Omitted group is fans with team eliminated from playoffs already entering study. Playoff Team is a dummy for if favorite team is in the playoffs entering the study. Chance Team is a dummy for if favorite team has a chance to make the playoffs entering the study.

TeamTV reports the number of team games the fan watched on TV in previous season. TeamLive reports the number of team games the fan watched live in previous season. AllLive reports the number of NFL games the fan watched live in previous season.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.10: *NFL Fans Study: Principal Components of Emotion*

	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6
Eigenvalues	3.424368	1.226232	.5973393	.3114762	.2604462	.1801385
Eigenvectors						
Excited	-.4039463	.4142351	.4786554	.1691208	.3808856	-.5122988
Happy	-.4703287	.2139072	.3553581	-.0119421	-.1429311	.7656261
Sad	.4460407	.247348	.2917086	.6615501	-.465911	-.0071545
Nervous	.0070041	.8024933	-.5903144	-.0655855	-.0307083	.0473282
Angry	.4268865	.2670605	.451166	-.7121362	-.1772166	-.0659707
Disappointed	.4841332	.0784554	.0865675	.1488877	.764902	.3804251

This table presents the results of a principal components analysis of the relationship between current happiness, sadness anger, disappointment, nervousness, and excitement.

This table is based on both baseline and post-game vectors of emotions for each participant.

The total sample size is 442 observations.

Table C.11: *NFL Fans Study: Effect of Team Win or Loss on Specific Emotions*

	(1) Sad	(2) Sad	(3) Angry	(4) Angry	(5) Disappointed	(6) Disappointed	(7) Happy	(8) Happy
Team Won	-6.97***	-12.7	-4.25**	1.31	-27.6***	-20.8**	29.5***	41.3***
	(1.83)	(18.8)	(2.03)	(21.6)	(2.97)	(9.09)	(3.14)	(14.0)
Team Lost	43.0***	39.2**	39.9***	46.5**	47.2***	57.7***	-43.9***	-36.7***
	(3.58)	(18.5)	(3.68)	(21.0)	(3.30)	(8.05)	(2.85)	(13.1)
Constant	11.2***	24.7***	10.4***	36.7***	36.0***	65.6***	51.3***	59.7***
	(1.61)	(7.66)	(1.66)	(12.1)	(2.63)	(15.8)	(2.46)	(8.93)
Observations	442	442	442	442	442	442	442	442
R^2	0.447	0.743	0.370	0.715	0.575	0.787	0.569	0.759
Participant Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Week Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes

Standard errors in parentheses.

This table shows the results of a regression of the current emotion variables on game outcome (win/loss), with the baseline survey responses serving as the omitted comparison group.

Each emotion variable has two columns above - one reporting regression results without participant and week fixed effects, and one with these fixed effects. Note that the four variables in models 1-8 are self-reported current emotions on a 0-100 scale.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.12: *NFL Fans Study: Effect of Team Win on Risk Outcomes*

	(1) Total Pumps	(2) Total Pumps	(3) Adj. Pumps	(4) Adj. Pumps	(5) Risk Aversion	(6) Risk Aversion
Team Won	0.313	0.254	0.157	-0.0280	0.0398	0.0333
	(0.306)	(0.301)	(0.172)	(0.160)	(0.0358)	(0.0370)
Baseline	2.168***	0.863***	0.886***	0.133	0.215***	-0.0250
	(0.336)	(0.315)	(0.171)	(0.170)	(0.0309)	(0.0265)
Constant	4.887***	2.568***	1.754***	1.267**	-0.348***	0.0125
	(0.223)	(0.536)	(0.136)	(0.582)	(0.0286)	(0.0174)
Observations	442	442	352	352	442	442
R^2	0.107	0.711	0.091	0.767	0.127	0.745
Participant Fixed Effects	No	Yes	No	Yes	No	Yes
Week Fixed Effects	No	Yes	No	Yes	No	Yes

Standard errors in parentheses.

The coefficient on Team Won represents the effect relative to the Team Losing, which is the omitted group in these regressions.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.13: *NFL Fans Study: First-Stage IV Regression (Y=Negative Emotion)*

	(1) PCA Negative Emotion	(2) PCA Negative Emotion
Team Won	-44.51***	-51.36*
	(4.145)	(29.68)
Team Lost	92.59***	91.43***
	(4.749)	(29.02)
Constant	-13.83***	10.24
	(3.287)	(20.40)
Observations	442	442
R^2	0.682	0.839
Participant Fixed Effects	No	Yes
Week Fixed Effects	No	Yes

Standard errors in parentheses.

Table presents the first-stage IV Regression, regressing negative emotion on favorite team victory/loss.

The first model is without participant and week fixed effects, and the second model includes fixed effects.

The model with fixed effects is used in the second-stage analysis.

The negative emotion variable ranges in value from -88 to 137 in the data.

Baseline survey responses represent the omitted category (outcomes with 'no game').

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.14: *NFL Fans Study: Effect of Negative Emotion on Total Pumps (Naive and IV Second-Stage)*

	(1) Naive	(2) Naive w/ FEs	(3) IV	(4) IV w/ FEs
Negative Emotion	-0.00574***	-0.00204	-0.00632***	-0.00178
	(0.00191)	(0.00178)	(0.00214)	(0.00211)
Constant	5.776***	3.464***	5.775***	3.450***
	(0.140)	(0.565)	(0.140)	(0.540)
Observations	442	442	442	442
R^2	0.016	0.711	0.017	0.711
Participant Fixed Effects	No	Yes	No	Yes
Week Fixed Effects	No	Yes	No	Yes

Standard errors in parentheses.

Table presents the results of regressions of total pumps on negative emotion, both in 'naive' form and with IV methods used.

Specifications 1-2 present 'naive' results, without using game outcome as an instrument for emotion, both with and without fixed effects.

Specifications 3-4 present results using game outcome as an instrument for negative emotions.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.15: *NFL Fans Study: Effect of Negative Emotion on Adjusted Pumps (Naive and IV Second-Stage)*

	(1) Naive	(2) Naive w/ FEs	(3) IV	(4) IV w/ FEs
Negative Emotion	-0.00232**	0.000387	-0.00255**	0.000196
	(0.00107)	(0.000962)	(0.00117)	(0.00112)
Constant	2.106***	1.388**	2.111***	1.398**
	(0.0691)	(0.582)	(0.0687)	(0.583)
Observations	352	352	352	352
R^2	0.014	0.768	0.015	0.767
Participant Fixed Effects	No	Yes	No	Yes
Week Fixed Effects	No	Yes	No	Yes

Standard errors in parentheses.

Table presents the results of regressions of adjusted pumps on negative emotion, both in 'naive' form and with IV methods used.

Specifications 1-2 present 'naive' results, without using game outcome as an instrument for emotion, both with and without fixed effects.

Specifications 3-4 present results using game outcome as an instrument for negative emotions.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.16: *NFL Fans Study: Effect of Negative Emotion on Risk Aversion (Naive and IV Second-Stage)*

	(1) Naive	(2) Naive w/ FEs	(3) IV	(4) IV w/ FEs
Negative Emotion	-0.000413*	-0.000182	-0.000485**	-0.000233
	(0.000232)	(0.000239)	(0.000239)	(0.000259)
Constant	-0.256***	-0.0129	-0.256***	-0.0101
	(0.0128)	(0.0224)	(0.0128)	(0.0191)
Observations	442	442	442	442
R^2	0.011	0.745	0.012	0.745
Participant Fixed Effects	No	Yes	No	Yes
Week Fixed Effects	No	Yes	No	Yes

Standard errors in parentheses.

Table presents the results of regressions of risk aversion on negative emotion, both in 'naive' form and with IV methods used.

Specifications 1-2 present 'naive' results, without using game outcome as an instrument for emotion, both with and without fixed effects.

Specifications 3-4 present results using game outcome as an instrument for negative emotions.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$