EXPERIMENTS AND EXTERNALITIES:
UNDERSTANDING CAUSE AND EFFECT IN ENVIRONMENTAL DECISION MAKING

Greer Kathryn Gosnell

September 2016

A thesis submitted to the Department of Geography & Environment of the London School of Economics and Political Science for the degree of Doctor of Philosophy in Environmental Economics

DECLARATION

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

I declare that this thesis consists of 39,140 words.

STATEMENT OF CONJOINT WORK

I confirm that Chapter 3 was jointly co-authored with Professor John List and Dr. Robert Metcalfe at the University of Chicago, and I contributed 75% of this work. Specifically, I was responsible for the design and implementation of the field experiment—as well as the debrief survey that forms the backbone of Chapter 4—data analysis, and a vast majority of the writing and editing of the paper. I played a smaller role in designing the theoretical model, which features in the Appendix. To make the published version of Chapter 3 self-contained, it incorporates small sections that appear in Chapter 4 of this thesis, which is single-authored by the candidate.

I confirm also that Chapter 5 was jointly co-authored with Dr. Alessandro Tavoni of the Grantham Research Institute at LSE, and I contributed 50% of this work. We jointly made experimental design decisions, after which I coded the experiment and implemented it in the lab with Alessandro. I was responsible for conducting the data analysis, and we played equal roles in writing the paper.

STATEMENT OF PRIOR PUBLICATION

A version of Chapter 3 is publicly available as National Bureau of Economic Research Working Paper No. 22316 (Gosnell, List, and Metcalfe, 2016). A brief synopsis of the material is published in the Harvard Business Review (https://hbr.org/2016/08/virgin-atlantic-tested-3-ways-to-change-employee-behavior).

ABSTRACT

The field of behavioral economics enhances the ability of social science research to effectively inform socially efficient climate policy at the microeconomic level, in part due to the dependence of climate outcomes upon present and future human consumption patterns. Since the behavioral field is relatively new, environmental and resource economists still have scarce evidence as to why people make particular decisions. For this thesis, I have conducted both field and laboratory experiments to address market failures highly relevant to environmental outcomes, namely international public goods problems and externalities from fuel and resource consumption.

My methodology capitalizes upon the benefits of each experimental methodology—laboratory, artefactual, framed, and natural—to capture the effects of particular informational and contextual elements on subsequent behavior. While each methodology has its potential advantages and shortcomings, I contend that the complete toolkit is necessary to study a broad range of relevant environmental contexts. For instance, while natural field experiments are generally considered the "gold standard" in terms of exogeneity and generalizability, many settings in which field experimentation may provide tremendous insight preclude randomization across unknowing subjects. Similarly, researchers may not have access to populations of interest, though lab experimentation may still provide insights into the behavior of these populations or reveal motivations not yet captured in neoclassical utility functions. In this thesis, I will detail results from one of each experimental type, each suited to the context of interest.

The natural field experiment in Chapter 2 aims to discern whether there is a role for environmental preferences and cognitive dissonance to play in encouraging individuals to engage in resource-conserving behaviors, and suggests that the latter may be effective in changing the behavior of green consumers. Chapter 3 presents the results of a large-scale framed field experiment comprising all eligible captains in Virgin Atlantic Airways, which tested the impacts of personalized information, tailored targets, and prosocial incentives on captains' performance of fuel-efficient behaviors. In addition to documenting a substantial Hawthorne effect, we provide intent-to-treat estimates of the three types of feedback to show that tailored targets are the most (cost) effective strategy of those implemented. I introduce a complementary artefactual field experiment in Chapter 4, which allows for detailed scrutiny of captains' fuel efficiency based on their social preferences as well as preferences and attitudes toward risk and uncertainty.

I find that more risk-averse captains are more prone to over-fuel, that prosocial incentives increase captains' well-being, and that revealed altruism increases responsiveness to prosocial incentives. Finally, Chapter 5 aims to provide insight into the effects of "side deals" in facilitating cooperation on international climate agreements. Using a lab experiment, we find that side deals alter the composition of group contribution to climate change mitigation, eliciting increased effort on the part of players with higher wealth.

ACKNOWLEDGMENTS

Comprehensive acknowledgment of those who have played integral roles in helping us to achieve our aspirations is a crucially important yet impossible feat. To the many hundreds of individuals who have provided logistical, intellectual, or emotional support—including those not explicitly mentioned—I express the most sincere gratitude.

I am entirely grateful for the time, co-authorship, and counsel of my supervisor Alessandro Tavoni, who has made my doctoral pursuit both enjoyable and challenging. I also express gratitude to Carmen Marchiori, whose supervision in the first two years of my Ph.D. was invaluable. Additionally, I thank my co-authors John List and Robert Metcalfe for their trust, guidance, and contribution to the truly novel research that I hope will open doors for the environmental economists of my generation.

I owe great thanks to the Grantham Foundation who sponsored my studentship at LSE, and to the Grantham Research Institute on Climate Change and the Environment, whose devoted directors, researchers, and staff continually provide an environment supportive of curiosity and research innovation.

This thesis would not have been possible without the support and time of several key individuals within partner organizations. I give special thanks to Emma Harvey, Claire Lambert, Dave Kistruck, and Paul Morris at Virgin Atlantic Airways for their enthusiasm and exceptionally hard work to facilitate and implement the framed and artefactual field experiments that comprise half of this thesis. Personnel at OSyS, Qualtrics, and MasterCard also played integral roles. Neither study would have been possible without the generous funds of the Templeton Foundation.

Additionally, I give my sincerest thanks to Juliet Davenport, Dave Ford, Will Vooght, and other team members at Good Energy, whose passionate environmental stewardship are unsurpassable and whose competence allowed for the successful implementation of the natural field experiment featured in this thesis.

I cannot neglect to express the most genuine gratitude to those mentors whose guidance and belief in my capabilities directed me toward this path, especially Dr. Dorothea Herreiner, Dr.

Andrew Healy, and Dr. Brian Treanor. Your genuine interest in both the development and achievement of your students' passions and pursuits is truly rare, and I am beyond fortunate to have been under your wings at Loyola Marymount University.

And last, but certainly not least, I thank my beloved family and friends for their undying support through the highs and lows of this unusual journey. In moments of breakthrough and achievement, you celebrated my successes as if they were your own. In moments of depleted inspiration and personal hardship, you reminded me of my dreams and expressed your sincere faith in my potential. You have somehow convinced me that no goal is unreachable. To you I owe everything.

TABLE OF CONTENTS

CHAPTER I

EXPERIMENTAL FOUNDATIONS IN ENVIRONMENTAL ECONOMICS

"Three elements of policy for mitigation are essential: a carbon price, technology policy, and the removal of barriers to behavioural change. Leaving out any one of these elements will significantly increase the costs of action."

-Stern Review (2007)

1. Introduction

The global climatic effects of anthropogenic greenhouse gas (GHG) emissions arguably constitute the most pressing political and economic issue of the 21st century. As climate science becomes less uncertain regarding human impacts on the global environment, the appropriate response of industry and politicians is increasingly polarizing (e.g., Stern, 2007; Nordhaus, 2007). The uncertain and potentially extreme long-term consequences of climate change undoubtedly validate calls for coordinated political and economic action on a grand scale, while the short-term incentives of politicians and polluting industries—together with information deficiencies and limited state funds—preclude straightforward policy solutions (Stavins, 1998; List and Sturm, 2006; Allcott and Mullainathan, 2010). Market-based mechanisms and regulation are politically unpopular, and their implementation requires lengthy processes that can result in failure to realize intended objectives (e.g., Jaffe and Stavins, 1995; Convery, 2009). As nations seek solutions on national and international scales, growing atmospheric GHG contents continue to alter essential ecosystems while increasing the unpredictability of future climate outcomes. A crucial and immediate objective of policymakers, therefore, is to reveal and implement climate change initiatives that are simultaneously politically feasible and cost-effective.

Despite historical political failure to curb emissions on a coordinated global scale, small-scale actors—including private businesses and local governments—have taken action independently. In all sectors, there are low- and negative-cost options for reducing GHG emissions (Enkvist, Nauclér, and Rosander, 2007; Stern, 2007). That is, reducing emissions is actually efficiency *enhancing*, but hidden costs (e.g., high transaction costs and lack of salient

information) preclude the uptake of such options. Many of these 'low-hanging fruits' rest upon a change in private consumption or technology adoption behavior. Overcoming these hidden costs is essential to realizing immediate climate benefits.

A growing body of experimental literature attempts to break down these unseen barriers by determining the drivers of prosocial human behavior (e.g., List and Lucking-Reiley, 2002; Johnson and Goldstein, 2003; Karlan and List, 2007; Meier, 2007; Ariely, Bracha, and Meier, 2009). Studies in the environmental realm have, for example, demonstrated the effects of social norms on water consumption (Ferraro and Price, 2013) and residential energy use (e.g., Allcott, 2011a; Dolan and Metcalfe, Mimeo), as well as the impact of defaults on carbon offsetting behaviors (Löfgren et al., 2012; Araña and León, 2013).

Policymakers are increasingly relying upon such experimental research in the social sciences for policy and project decision making. Recent findings of behavioral economics indicate a growing need to enhance both the quality and quantity of empirical research in policy-relevant realms and to incorporate the findings into decisions on government regulation (Sunstein, 2011). Empirical research to date already suggests that there are a number of existing inefficiencies due to government and market failures—e.g., lack of information, externalities, and public goods—that inexpensive policy measures can help to surmount (for examples, see DellaVigna, 2009).

"Libertarian paternalism" in policymaking may work to overcome many of the political economic concerns associated with government intervention to correct market failure, if applied ethically and cautiously (Glaeser, 2006; Thaler and Sunstein, 2008; Reisch and Sunstein, 2014). Proposals for governments' adoption of this concept derive from its potential to breed bipartisan (or multi-partisan) agreement, in that libertarian policies do not restrict citizens' choices but rather capitalize upon behavioral tendencies in order to enhance social well-being in ways that are empirically demonstrated to be effective. In this thesis, I will discuss various interventions into relevant decision-making contexts that aim to decrease the extent of globally harmful environmental externalities. I will focus on the extent to which decision contexts, individual preferences (e.g., altruism, risk), and behavioral 'biases' can inform micro-level climate change mitigation policy at the government and industry levels. I argue that policymakers and industry leaders should, where possible and cost-effective, use economic experimentation—the foremost methodological standard in economics for making causal inference—as part of a comprehensive

environmental policy strategy. Specifically, I aim to demonstrate the benefits of both lab and field experimentation in understanding human behavior with respect to global climate agreements as well as resource (i.e. paper and fuel) use.

This thesis is structured as follows. This chapter will provide background on the shortcomings of neoclassical economics and the consequent incorporation of concepts and findings from psychology and experimental economics into the burgeoning field of behavioral economics, concluding with short descriptions of the papers to follow. The remainder of the thesis comprises four experiment-based essays—inclusive of each experimental method, as classified by Harrison and List (2004)—and a brief concluding chapter. Chapter 2 details the implementation and results of a natural field experiment with a UK utility that uses interventions rooted in psychology to reduce paper use via consumer switching to online billing. Chapter 3 introduces a large-scale framed field experiment in the aviation industry aimed at reducing the contingency fuel use of airline captains. Chapter 4 extends this research into the well-established economic domains of agency theory and risk preference elicitation, providing novel insights into the effects of captains' preferences and attitudes toward risk and uncertainty, as well as their social preferences, on performance outcomes. Chapter 5 demonstrates the potential for lab experimentation to inform global diplomatic structures within which national governments may enter into binding climate change mitigation agreements. Finally, the conclusion summarizes my findings and provides implications for researchers, policymakers, and industry practitioners.

2. The decline of *Homo economicus*

Traditional economic theories of decision making are contingent upon neoclassical assumptions of human rationality and optimization. Neoclassical economic theory presumes that human incentives are purely self-interested and unboundedly rational (DellaVigna, 2009). That is, human beings are assumed to predicate all decisions on an internal calculation of their own self-serving and well-behaved utility functions. This purely rational and optimizing creature is commonly referred to as *Homo economicus* ("economic man"), and economists generally assign to him four basic features: 1) well-defined preferences entering into a utility function that he seeks to maximize; 2) preferences accurately reflecting the true costs and benefits of the available alternatives; 3) developed beliefs that inform him as to how uncertainty resolves itself; and 4) the ability to update beliefs when new information becomes available (Camerer et al., 2003).

Over the last few decades, the neoclassical interpretation of utility maximization has been critiqued for its inability to explain particular real-world behavioral phenomena.[1] Behavioral economics has consequently emerged from the convergence of psychological and economic studies. The field initiated with the publication of empirical findings on reference dependence and loss aversion (or "the endowment effect"). The endowment effect was evident in a simple experiment that tested the willingness-to-pay and willingness-to-accept of experimental subjects for mugs randomly distributed to half of the study's participants (Kahneman, Knetsch, and Thaler, 1990, 1991). The median selling price was $5.79, while the median buying price was only $2.25, less than half of the former. To control for wealth effects, they reported another study in which "buyers" became "choosers" and were asked whether they would prefer to take the money or the mug for a series of monetary amounts. Still, the median selling price was twice the median choosing price. Kahneman and Tversky's (1979) Prospect Theory, which seeks to model 'real'—as opposed to 'optimal'—human decision making under conditions of risk and uncertainty, is often used to explain this and other anomalistic findings indicating reference dependence (Abeler et al., 2011; Farber, 2008).[2]

Kahneman and Tversky also contend that, under conditions of uncertainty, humans are driven by heuristic shortcuts that generate probability judgments straying from statistical reality. Standard economic models assume that humans make probability judgments based on mathematical calculations derived from prior experience; for instance, Bayes' rule states that humans statistically update their probability judgments when confronted with new information. In reality, heuristic "shortcuts" drive a wedge between subjective and objective probabilities, so that standard models will not be descriptive of behavior (Tversky and Kahneman, 1973).

In sum, the founding of Prospect Theory uncoupled the normative framework of decision making—that of rational utility maximization—from actual observed behavior, deeming the normative model incompatible with real world outcomes. Their findings made clear the benefits—even the necessity—of incorporating psychological approaches into the economic

[1] The recent emphasis of social scientists—and particularly economists—on the importance of empirical results to supplement theory stems from the increasingly clear distinction between *Homo economicus* (or "rational economic man") and *Homo sapien*. That is, actual human behavior has been shown to deviate markedly in several respects from the predicted behavior of the neoclassical economic agent. For instance, human responses to stimuli may depend upon the manner in which decisions are presented, how others behave, or initial reference points (for overviews, see Dawnay & Shaw, 2011; Dolan et al., 2012).

[2] Kahneman and Tversky (1979) uncovered the irrational effects of risk aversion and framing of alternatives—which can alter perceptions of true value—on the decision-making process. They find a "certainty effect," a higher aversion to reductions in probability from a reference point of attaining an outcome with certainty than to reductions in probability from a reference point where probability is lower, even though expected utilities remain the same.

discipline and considerably challenged traditional economic assumptions of rationality and self-interest.

Needless to say, these early discoveries spurred the inception and evolution of behavioral economics and, inevitably, initiated a reversion of *Homo economicus* to *Homo sapien* in many policy-relevant contexts. Non-traditional economists have since unveiled several additional irrational decision-making tendencies. For example, humans can be strong reciprocators; that is, they may respond to free-riding actors through retaliation, even if there does not exist a realistic expectation of personal gain or if there is a personal cost (Gintis, 2000; Fehr and Gächter, 2000). They tend to discount the future hyperbolically, employing a higher discount rate in the near term than in the more distant future (Laibson, 1997; Ashraf, Karlan, and Yin, 2006; Hepburn, Duncan, and Papachristodoulou, 2010). These and other deviations from neoclassical theory have been termed "behavioral failures" (Shogren and Taylor, 2008), also often described in the literature as anomalies, heuristics, paradoxes, and biases. Mullainathan and Thaler (2000) group these failures into three categories: bounded rationality (e.g., endowment effects, heuristics, and self-serving bias), bounded self-interest (e.g., other-regarding preferences), and bounded willpower (e.g., lack of self-control). As a complementary means to traditional policy solutions that aim to overcome market failure, this thesis focuses on the potential to capitalize upon bounded rationality and bounded self-interest to increase environmental efficiency and welfare.

3. Behavioral failure: Are markets to blame?

While behavioral failure appears to permeate through many economic sectors, neglecting to acknowledge its consequences may be especially toxic for policies targeting environmental sectors. Traditional economists point to particular market failures—such as missing markets, non-excludable/non-rival goods, and imperfect information—to explain social and environmental inefficiencies. In the absence or weakness of markets for most environmental amenities, individuals act as "asocial beings," unaccountable to others due to the inability of collective market institutions to arise (Crocker, Shogren, and Turner, 1998). While studies have shown that individuals behave more rationally with experience (List, 2003), ecosystem services often lack markets within which experience may be accumulated. These market-focused claims suggest that creation of markets for environmental goods and services will lead to environmental efficiency.

Accordingly, the notion that market failure is the sole problem underpinning environmental damage has historically driven most policy research and outcomes. Tools such as payment for ecosystem services and carbon pricing (i.e. emissions permits and carbon taxes) rest upon the assumption that human beings are (or behave as though they are) rational actors (Shogren and Taylor, 2008). While market failures undoubtedly play a prominent role in environmental inefficiency, this assumption is clearly unrealistic and suggests that market failures cannot fully explain environmental deadweight loss. Is it possible to lessen the severity of these commonly cited market failures in the environmental sector—especially in those contexts where markets are unlikely to succeed or exist—by targeting human behavior? Is market failure the lone culprit, or does behavioral failure exacerbate environmental inefficiencies? Incorporation of behavioral anomalies into economic models may provide potentially significant scope for more efficacious and cost-effective environmental policy (Kunreuther and Weber, 2014).

Behavioral economics has already shaped research in areas within environmental economics involving risk, valuation, and strategic interaction (Shogren, 2012). Perhaps the most prominent example followed the Exxon Valdez oil spill, when a NOAA panel debated—and ultimately upheld—the merits of using stated preference methods (SPMs) to conduct a valuation of social damage, thereby identifying a comprehensive social compensation value. The disparity between willingness-to-pay (WTP) and willingness-to-accept (WTA; Horowitz and McConnell, 2002), which follows from behavioral failures such as the endowment effect and status quo bias, led to controversy and subsequent alterations in non-market valuation using SPMs. The panel held similar discussions around an embedding effect (Kahneman and Knetsch, 1992) or insensitivity to scope (Desvousges, 1992), claiming that this behavioral failure was the "most important internal argument" against SPM valuation (Arrow et al., 1993). Explanations for scope insensitivity may rest upon other behavioral aspects that are rarely acknowledged in rational choice theory, such as warm glow effects and mental accounting (see Andreoni, 1989; Thaler, 1999). While this debate remains controversial, there have been many subsequent applications of behavioral economics to environmental contexts (for a comprehensive overview see Shogren, 2012). Given that markets for environmental goods and services are difficult to establish, such application is widely viewed as a promising step toward the enhancement of environmental policy and welfare.

4. A role for experimentation

While economists have developed a large suite of tools for identifying correlations and inferring causation under various assumptions (Angrist and Pischke, 2010), studies of observational data that aim to infer causality often suffer from critiques related to endogeneity. A seminal paper by Vernon L. Smith argued for a transformation of the economic discipline from a social science based on observational data to an experimental science rooted in exogeneity (Smith, 1982). Randomization has swiftly become one of the most embedded and important tools in the economist's toolkit, and Smith—along with psychologist and behavioral economist Daniel Kahneman—was recognized for his contribution to the discipline with the Nobel Memorial Prize in Economic Sciences in 2002. The simultaneous awarding of this honor to pioneers in experimental and behavioral economics is testament to the two fields' complementarity and symbiotic nature; each field acts as a support and a catalyst for the other. For example, the application of this ongoing relationship to contingent valuation has improved insights and methodologies surrounding non-market valuation, a concept with extreme relevance to environmental economists (List and Price, 2013).

Yet, lab experiments often suffer criticisms associated with generalizability. Critics question whether student behavior can inform our understanding of economic agents in the field, and whether variance in the size or type of incentive offered render lab results inapplicable to many field contexts. To bridge the gap between randomization (a prominent tool for assessing causality in microeconomic systems) and observational data (a means to understand economic agents in relevant contexts), economists have more recently introduced field experimental methodologies (see Harrison and List, 2004). For instance, to eliminate the concern of sample generalizability, one may simply perform an "artefactual field experiment" (AFE) by implementing a lab experiment with a population of interest.

Better yet, one may further improve generalizability by executing the experiment in the context of interest, matching the field setting with respect to the scale and type of incentives as well as the specific behavior of interest. Oftentimes field experimentalists must operate within the constraints of businesses and governments or under strict requirements of ethics committees, and a common stipulation is that the researcher obtains subjects' consent, introducing potential for experimenter effects. When subjects are aware of their participation, the methodology is defined as a "framed field experiment" (FFE), according to the typology in Harrison and List

(2004). If subjects must participate voluntarily, the methodology may additionally suffer from selection bias. In contrast, a "natural field experiment" operates within the context of interest with the population in question, and subjects do not know their behavior is being observed. In this case, treatment effects are assumed to be causal and unbiased.

While the latter methodology may suffer potential flaws of its own—perhaps including generalizability to *other* field contexts and diminished control over the experimental environment (see Heckman and Smith, 1995; Deaton, 2010; and Al-Ubaydli and List, 2015, for a glimpse into these debates)—it is nonetheless among the cleanest methods for reaching the empirical gold standard of causal inference (List, 2011a). Of course, many questions pertinent to economics do not lend themselves to field experimental settings, and in these settings lab experiments can help to generate understanding of particular scenarios, perhaps especially when the researcher introduces contextual features relevant to the underlying research question at hand.

i. Laboratory experiments in environmental decision making

Lab experiments have long been used to identify and address environmentally relevant market failure. One of the most heavily researched lab experimental games is that of the voluntary contribution mechanism (VCM), which models a social dilemma mirroring the predicted "tragedy of the commons" scenario inherent in public goods (Hardin, 1968). Similar to findings from ultimatum and dictator games (see Güth, Schmittberger, and Schwartze, 1982; Camerer and Thaler, 1995), VCM experiments have established that individuals tend to display some degree of cooperation inconsistent with either the privately or socially optimal outcomes. While a vast amount of research has been done to identify frameworks and mechanisms that may lead to cooperative outcomes in public goods games (for a review see Ledyard, 1995), most of these experiments lack contextual features that make them directly relevant to environmental settings. That is, they tend to occur in computer laboratories on university campuses amongst college students, and they generally study abstract games involving tokens or money without relevance to particular public good contexts.

Perhaps the most pressing public goods issue in the field of environmental economics is that of atmospheric greenhouse gas emissions, a problem with a top-down solution that requires coordination with little enforcement capability. Since there are high barriers to conducting experiments with the sample of interest in this context (i.e. high-level policymakers), lab

experimentation that maximizes contextual relevance is perhaps the closest that researchers can get to causal inference in this particular domain. Thus, where field experimentation is impractical or infeasible, experimentalists may instead attempt to glean insights from well-designed lab experimentation.

The question of how applicable findings from VCM experiments in the lab may be to questions pertaining to global climate negotiations rests upon a number of contextual features. For instance, it is generally accepted that climate change requires a minimum amount of cooperation in order for *any* cooperation to be meaningful (i.e. it is a *threshold* public good; Tavoni and Levin, 2014; Dannenberg et al., 2014). Additionally, meaningful cooperation involves a large number of heterogeneous players (Kolstad, 2010; Barrett, 2010) who may exhibit self-serving bias in their perspectives on fairness (Babcock and Loewenstein, 1997; Babcock, Loewenstein, Issacharoff, and Camerer, 1995).

Therefore, a number of recent laboratory experiments have attempted to improve the generalizability of the standard VCM experiment by contextualizing the game to mimic the circumstances of global climate change negotiations (e.g., Milinski et al., 2008; Tavoni et al., 2011; Barrett and Dannenberg, 2013; Dannenberg et al., 2014). These games have suggested, for example, that group members only cooperate when the risk of loss is sufficiently high (Milinski et al., 2008), that inequality in wealth further complicates the coordination problem and pledges can aid in coordination (Tavoni et al., 2011), and that reducing uncertainty drastically facilitates cooperation (Barrett and Dannenberg, 2013; Dannenberg et al., 2014). The findings from such experiments foster discussion surrounding, for example, the importance of resolving uncertainties and amending the infrastructure within which global climate negotiations transpire.

ii. Field experiments in nonmarket valuation

In the case of missing markets for environmental amenities, the question of how to appropriately engage in non-market valuation is immensely important. Accordingly, the NOAA panel spurred a thread of literature examining the WTP-WTA gap, and experimental economic literature in this domain has since multiplied. Lab experiments in cheap talk (Cummings Harrison, and Osborne, 1995; Cummings and Taylor, 1999)—the explicit incorporation of hypothetical bias into the design of the contingent valuation survey—inspired experimental investigation into improvements in stated preference methodology. For instance, List (2001) and

Carlsson, Frykblom, and Lagerkvist (2005) implemented FFEs—the former among sports card traders and the latter among Swedish meat and poultry consumers—to investigate the impact of cheap talk, finding evidence in support of its effectiveness in lowering valuations. In the vein of List (2003) and List (2011b), several cheap talk FFEs have found that cheap talk influences behavior among subjects unfamiliar with the good being valued, while it does not influence experienced consumers of the good in question (e.g., List, 2001; Lusk, 2003; Aadland and Caplan, 2003, 2006; Blumenschein et al., 2008). Following from a FFE demonstrating the effects of consequentiality on contingent valuation responses regarding a referendum vote (Cummings and Taylor, 1998), Landry and List (2007) compare cheap talk and consequentiality and find that the results under the two scenarios are indistinguishable.[3]

An additional set of experiments explores the extent to which the context of survey administration influences valuation. List et al. (2004) use a FFE to study the effects of respondent anonymity, finding that social pressure plays a sizable role in valuation. In a natural field experiment (NFE), Alpizar, Carlsson, and Johansson-Stenman (2008) vary anonymity to study the impacts of donations to a Costa Rican national park, finding that social anonymity decreases donations by 25% relative to donations solicited by an interviewer. Leggett et al. (2003) conduct a similar experiment to detect differences in valuation responses for Fort Sumter National Monument, finding that self-administration of surveys leads to 23-29% lower valuations than do interview responses.[4] Clearly, lab and field experimentation have contributed greatly to our understanding of biases in contingent valuation—often considered the "only game in town" when it comes to total valuation of environmental amenities—encouraging the continuous evolution and refinement of the method to increase its credibility.

iii. Field experiments in waste disposal

Additionally, a number of small-scale experiments have contributed to our understanding of human behavior in the context of waste disposal. For example, experimental literature on littering highlights the tendency of individuals to litter significantly more when they are subject to littered environments as opposed to uncontaminated ones (Geller, Witmer, and Tuso, 1977;

[3] A number of other experiments explore the results and nuances of consequentiality. For instance, Herriges et al. (2010) implement a natural field experiment to value water quality in Iowa lakes, finding that the degree of consequentiality is unimportant in contingent valuation; so long as there is at least a weak perception of policy influence, valuations are consistent. Additionally, Vossler et al. (2012) find similar results in a repeated valuation study centered upon riparian buffers in Quebec.

[4] Additional experiments seek to reveal the nuances behind such experimenter effects, such as the formality of the interviewer (Bateman and Mawby, 2004), the association between the interviewer and the product (Loureiro and Lotade, 2005), and demographic characteristics of the interviewer (Gong and Aadland, 2011).

Krauss Freedman, and Whitcup, 1978; Reiter and Samuel, 1980; Cialdini, Reno, and Kallgren, 1990; Ramos and Torgler, 2010; Keizer, Lindenberg, and Steg, 2011).[5] Most of these studies involved varying the extent of litter or other contamination (e.g., graffiti) then distributing flyers and recording whether they were properly discarded. Dur and Vollaard (2012) implement a field experiment to test if policy efforts to erase this normative effect of littering—i.e. provision of public cleanup services—contribute to *less* littering (i.e. due to the "broken window effect") or *more* littering (i.e. due to free riding). Having randomized the frequency of cleanup services over three months, they find evidence for both phenomena, with free riding outweighing contributions.

Similarly, field experiments have revealed cost-effective means to increase responsible recycling behavior. These experiments have revealed the influence of exogenous goal setting (McCaul and Kopp, 1982), community leaders (Hopper and Nielsen, 1991), salience (Krendl, Olson, and Burke, 1992), and social norms (Schultz, 1999) on household-level recycling behavior. In a work environment, Holland, Aarts, and Langendam (2006) use a natural field experiment to demonstrate a significant effect of implementation intentions (i.e. planning) on recycling behavior. Randomizing group-level public exposure in a recent framed field experimental threshold public goods game, Alpizar and Gsottbauer (2015) capitalize upon human pride and (especially) aversion to guilt to promote increased recycling effort in Costa Rica. They additionally demonstrate a crowding-in effect of regulatory policy that reduces the risk of falling short of the threshold. Hence, while rational incentives are undoubtedly present in waste disposal and sorting decisions, policymakers may target a number of behavioral mechanisms to cost-effectively improve responsible disposal of waste.

iv. Field experiments in energy and water conservation

Relatively recently, field experimentation has been used to address resource conservation and attenuation of environmental externalities in the economy at large. For instance, several field experiments have demonstrated effectiveness of interventions in residential electricity and water use. Perhaps most famously, Opower's experiments demonstrate that provision of social norm information to households can lead to an average 2% savings in electricity use (Allcott, 2011b). Ayres, Raseman, and Shih (2012) also analyze Opower experiments, finding that the messages are most effective among households with the highest consumption and that the frequency of

[5] Few experimental studies reject this effect (e.g., Crump et al., 1977; Reno et al., 1993).

messaging matters.[6] Similarly, in a natural field experiment on residential water conservation, Ferraro and Price (2013) demonstrate that the use of normative messages is more effective than either prosocial appeals or technical information on their own, and again that high-use households—i.e. the most price insensitive subjects—are most susceptible to these messages. The norms' effects persisted well beyond the intervention relative to control groups (and relative to prosocial and technical messaging groups in the water conservation experiment), indicating that norms may 'nudge' individuals into making more energy- and water-efficient decisions on a habitual basis (Ferraro, Miranda, and Price, 2011; Allcott and Rogers, 2014). However, Tiefenbeck et al. (2013) warn against the potential negative spillovers, demonstrating a perverse effect of a water-saving campaign on residential electricity consumption.

In addition to testing norms-based messaging, experimental economists have explored the potential for dynamic pricing schemes to increase the efficiency of residential energy consumption. Wolak (2006) was the first to randomize dynamic energy pricing. In a field experiment in Anaheim, experimental subjects received smart meters and were assigned to remain on the business-as-usual pricing plan or to receive a critical peak pricing (CPP) plan. In the CPP plan, customers received rebates worth $0.35 per kWh reduction relative to their household's average use for the most consumptive non-CPP days during that time period. While treated customers reduced their consumption by 12% relative to control customers during critical peaks, there is a large perverse effect on non-CPP days due to the structure of the CPP plan, which incentivizes treated subjects to consume more on non-CPP days to increase their rebate. Additionally, customers were guaranteed their CPP bill would not exceed their bill under a standard increasing block tariff, potentially dampening incentive to reduce consumption.

In an extension, Wolak (2011) verifies the effectiveness of CPP plans in reducing consumption during peak events, especially when CPP does not simply reward a customer with rebates if consumption is below a reference level; the most effective treatment by far is the CPP plan where a customer pays the high tariff for every kWh consumed during peak events. Additionally, he investigates the existence of an individual "cost of taking action" phenomenon whereby an individual's cost of reducing energy use must be overcome by a sufficiently large price spike; he finds no evidence for such a cost of action, as (price adjusted) reductions in consumption on hourly and CPP tariffs are equal. Allcott (2011b) also detects significant effects of dynamic pricing in the context of salient hourly price changes, finding that consumers are

[6] Costa and Kahn (2013) detect heterogeneity in the effects of the home energy reports according to political ideology, demonstrating a perverse effect on households in conservative areas that is outweighed by the effect on households in liberal ones.

fairly price elastic, reducing consumption considerably during peak hours, and that they do not consequently increase consumption in off-peak hours.

Having demonstrated the effects of dynamic pricing, others have introduced interventions to evaluate additional or relative efficiencies. Jessoe and Rapson (2012) demonstrate an effect of real-time price change updates (0-7 percent), which grows considerably when interacted with real-time consumption feedback (8-22 percent), demonstrating the importance of salience in both price and quantity information. Kahn and Wolak (2013) also find that improved comprehension of marginal pricing schemes in combination with understanding of the consumption of electricity-consuming appliances reduces consumption 1.5 and 3 percent on average for the customer bases of two California utilities. Ito, Ida, and Tanaka (2015) compare dynamic pricing (i.e. extrinsic motivation) to moral suasion (i.e. intrinsic motivation). They find that consumption decreases monotonically with increases in the marginal CPP price; while moral suasion significantly reduces consumption during peak events (3 percent), the effect is only a fifth as strong as that of the marginal price increase (15 percent). Using follow-up data, they find that only economic incentives have a persistent effect, likely owing to habit formation and the incentives' effectiveness among low-income households.

Finally, a recent line of literature aims to understand and *price* the effects of behavioral phenomena on energy-saving technology adoption. For instance, in a door-to-door field experiment, Herberich, List, and Price (2011) structurally estimate the effects of social pressure and norms on the purchase of compact fluorescent light bulbs (CFLs), finding that both have an influence on buyers' decisions in this context. While social norms (i.e. informing the buyer of the proportion of similar households using CFLs) affect buyer decisions on the extensive margin—that is, *whether* to buy CFLs—price variation influences decisions on the intensive margin, so that buyers purchase more CFLs when they are cheaper. Interestingly, the data suggest that individuals who are not warned that a solicitor will approach the house may experience negative utility from the purchase due to social pressure to buy when confronted with the solicitor, a finding concordant with DellaVigna, List, and Malmendier (2012). Finally, Yoeli et al. (2013) demonstrate the importance of observability—what they term 'indirect reciprocity'—in voluntarily contributing to peak demand reductions.

In sum, economic experiments have revealed an impactful role for social norms and extrinsic incentives in promoting residential energy conservation. Further experimentation in this

domain will continue to reveal and refine mechanisms by which utilities and policymakers may influence consumption.

v. New directions in environmental field experimentation

Field experiments have come to play a ubiquitous role in development economics (Kremer, 2003; Banerjee and Duflo, 2006; Duflo, 2006; Glewwe and Kremer, 2006; Banerjee and Duflo, 2009). This thesis argues that environmental economists should capitalize upon the methodology in a similarly pervasive manner, creating exogeneity in relevant contexts to understand cause and effect in environmental decision making. This endeavor should be pursued along two dimensions: replication of well-documented observational and laboratory findings across a number of field contexts to determine external validity—as with Gneezy and List's (2006) field experimental test of the gift exchange findings in Fehr, Kirchsteiger, and Riedl's (1993) lab experiment, which endorse the use of efficiency wages—and pursuit of innovative insights regarding environmentally-relevant behavioral phenomena.

With regard to the former, some environmental economists have begun to investigate well-documented laboratory findings—or field findings from non-environmental contexts—in environmental field experiments. For instance, Stoop, Noussair, and van Soest (2012) explore the social dilemma problem of the VCM in the context of a privately owned fishery, replicating a VCM incentive structure in a framed field experiment among recreational fishermen. In contrast to the findings of myriad laboratory experiments, they find support for the classic theoretical outcome of no cooperation, refining the experimental design to determine that the reason for the divergence from lab outcomes is the nature of the activity. That is, contrary to the most common critique of laboratory experiments, the cause of the divergence is not the nature of the subject pool but rather simply that the fishermen enjoy the *act* of fishing. Other studies have replicated the well-established finding that defaults influence behavior in 'green' contexts such as reducing junk mail (Liebig and Rommel, 2014) and increasing uptake of carbon offsets (Löfgren et al., 2012), or the finding from the economics of charitable giving that 1:1 matching is most effective at inducing carbon offsetting contributions from bus travelers (Kesternich, Löschel, and Römer, 2014).

In addition to testing existing findings in environmentally relevant contexts, traditional and behavioral economists should target environmental behaviors to test bed new and existing

theories (see Card, DellaVigna, and Malmendier, 2011, for a discussion of the synergistic relationship between theories and experiments). For instance, theories of social norms have benefited from the randomization of norm information in a number of environmentally relevant contexts, significantly influencing policy on consumer behavior and opening up a large strand of literature in the energy space. As discussed below, this thesis presents an additional novel context for field experiments to reduce externalities by using interventions to target excess fuel use in the aviation industry. Such research will serve to demonstrate the benefits to practitioners of participation in academic research—thereby mainstreaming such mutually beneficial partnerships— and to open new and important lines of academic inquiry.

In sum, the field is rife with opportunity for creative and intellectual minds to identify behavior-induced environmental inefficiencies and test the effectiveness of various interventions in eliminating social deadweight loss, especially in contexts where Pigouvian taxes, permits, or regulation are impractical or infeasible in the short run. On a large scale, the methodology can quickly narrow in on socially and privately cost-effective means to accomplish the environmental goals of consumers, businesses, and policymakers alike.

5. Dissertation outline

This thesis aims to demonstrate a role for the various types of experimental methodology in furthering the field of environmental economics. The first paper (Chapter II) describes a simple natural field experiment across the customer base of Good Energy—a British all-renewable energy supplier—that aims to maximize take-up of its new online billing service. The goal of the study is to understand the effect on 'green' consumers of environmental information and cognitive dissonance by varying message content in a regular email communication with customers. In one treatment, we randomize messaging centered on the environmental benefits of switching to e-billing. In an alternative message, we remind consumers of their 'green' identity and past environmental decision making in signing up to Good Energy. We compare these treatments to a control email highlighting improved access to bills and customer satisfaction, and we find that environmental messaging does not enhance customers' proclivity to sign up to e-billing, while cognitive dissonance weakly improves take-up.

The second and third papers investigate means by which to reduce greenhouse gas emissions from the aviation sector by implementing experimental treatments rooted in a novel

economic theory of captain behavior (Chapter III) and by exploring whether captains' preferences and attitudes toward risk and uncertainty—as well as their social preferences—play a role in their fuel use (Chapter IV). From February 2014 through September 2014, all eligible captains in Virgin Atlantic Airways knowingly participated in a field experiment to test whether performance feedback, tailored performance targets, and prosocial incentives influence captains' implementation of three specific fuel-efficient behaviors. The advantage of this FFE relative to most others is that all eligible captains were included in the subject pool (i.e. the airline opted to include the entire population, not just a subset of volunteers), so the results are free from selection bias. We find that exogenous performance targets are the most cost-effective means by which to improve fuel efficiency; we show in the following paper that prosocial incentives—which are equally effective in terms of changing behavior—improve captains' job satisfaction. Given Kahneman and Tversky's (1981; p. 211) perspective that "the adoption of a decision frame is an ethically significant act" in contexts where it "influences the experience of consequences," these findings leave airlines with both efficiency and experiential perspectives to consider.

We additionally implemented an AFE in the study debrief survey in which we ask incentive-compatible risk preference elicitation questions, evaluate risk attitudes, and assess uncertainty aversion. In Chapter IV, we investigate whether these innate heterogeneities play a role in captains' fueling decisions, finding that increased risk tolerance on the self-reported attitudinal risk scale improves efficiency in aircraft fuel loading decisions. In addition, among captains who receive prosocial incentives in the above framed field experiment, captains who donated more to charity prior to the study perform more strongly once the experiment begins than do captains who donated less. As mentioned previously, captains who receive prosocial incentives have higher job satisfaction than those captains in the control group, indicating that captains have positive levels of altruism that may be harnessed for the improvement of social outcomes as well as employee satisfaction.

The final paper (Chapter V) describes a novel Nash bargaining experiment in the laboratory with groups of six heterogeneous countries bargaining over a fixed allocation of global emissions. The experiment aims to understand whether side deals—i.e. agreements among homogeneous subsets of countries that occur prior to entering a global negotiation—can improve prospects for reaching effective global agreement to allocate emissions reductions consistent with a collective target. Given that the population of interest here consists of heads of state and high-level climate delegates to the United Nations Framework Convention on Climate Change, the

research question precludes the design and implementation of a field experiment; thus, the appropriate experimental tools exist solely in the laboratory. We find that inserting pre-negotiation side agreements into the global bargaining infrastructure does not increase the likelihood of success unless there is a strong signal of commitment by relatively rich countries; however, the existence of side agreements does serve to reduce the demand of the rich in global negotiations.

REFERENCES

Aadland, David, and Arthur J. Caplan. 2003. "Willingness to pay for curbside recycling with detection and mitigation of hypothetical bias." *American Journal of Agricultural Economics* 85 (2): 492-502.

Aadland, David, and Arthur J. Caplan. 2006. "Cheap talk reconsidered: New evidence from CVM." *Journal of Economic Behavior and Organization* 60 (4): 562-578.

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. "Reference points and effort provision." *American Economic Review* 101 (2): 470-492.

Al-Ubaydli, Omar and John A. List. 2015. "Do natural field experiments afford researchers more or less control than laboratory experiments?" *American Economic Review: Papers & Proceedings* 105 (5): 462-466.

Allcott, Hunt. 2011a. "Social norms and energy conservation." *Journal of Public Economics* 95 (9): 1082-1095.

Allcott, Hunt. 2011b. "Rethinking real-time electricity pricing." *Resource and Energy Economics* 33: 820-842.

Allcott, Hunt, and Sendhil Mullainathan. 2010. "Behavior and energy policy." *Science* 327 (5970): 1204-1205.

Allcott, Hunt, and Todd Rogers. 2014. "The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation." *American Economic Review* 104 (10): 3003-3037.

Alpízar, Francisco, Fredrik Carlsson, and Olof Johansson-Stenman. 2008. "Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica." *Journal of Public Economics* 92 (5): 1047-1060.

Alpízar, Francisco, and Elisabeth Gsottbauer. 2015. "Reputation and household recycling practices: Field experiments in Costa Rica." *Ecological Economics* 120: 366-375.

Andreoni, James. 1989. "Giving with impure altruism: Applications to charity and Ricardian equivalence." *The Journal of Political Economy* 97 (6): 1447-1458.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2010. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of Economic Perspectives* 24 (2): 3-30.

Araña, Jorge E., and Carmelo J. León. 2013. "Can defaults save the climate? Evidence from a field experiment on carbon offsetting programs." *Environmental and Resource Economics* 54 (4): 613-626.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." *American Economic Review* 99 (1): 544-555.

Arrow, Kenneth J., Robert S. Solow, Edward Learner, Paul Portney, Ray Rodner, and Howard Schuman. 1993. "Report of the NOAA panel on contingent valuation." *Federal Register* 58 (10): 4601-4614.

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. "Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines." *Quarterly Journal of Economics* 121 (2): 635-672.

Ayres, Ian, Sophie Raseman, and Alice Shih. 2012. "Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage." *Journal of Law, Economics, and Organization* 29 (5): 992-1022.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased judgments of fairness in bargaining." *American Economic Review* 85 (5): 1337-1343.

Babcock, Linda, and George Loewenstein. 1997. "Explaining bargaining impasse: The role of self-serving biases." *Journal of Economic Perspectives* 11 (1): 109-126.

Banerjee, Abhijit, and Esther Duflo. 2006. "Addressing absence." *Journal of Economic Perspectives* 20 (1): 117-132.

Banerjee, Abhijit, and Esther Duflo. 2009. "The experimental approach to development economics." *Annual Review of Economics* 1: 151-178.

Barrett, Scott. 2010. "Comment on 'Equity, heterogeneity and international environmental agreements'." *The B.E. Journal of Economic Analysis and Policy* 10 (2): 1-4.

Barrett, Scott, and Astrid Dannenberg. 2013. "Sensitivity of collective action to uncertainty about climate tipping points." *Nature Climate Change* 4: 36-39.

Bateman, Ian J., and James Mawby. 2004. "First impressions count: Interviewer appearance and information effects in stated preference studies." *Ecological Economics* 49: 46-55.

Blumenschein, Karen, Glenn C. Blomquist, Magnus Johanneson, Nancy Horn, and Patricia Freeman. 2008. "Eliciting willingness to pay without bias: Evidence from a field experiment." *The Economic Journal* 118: 114-137.

Camerer Colin, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue, and Matthew Rabin. 2003. "Regulation for conservatives: Behavioral economics and the case for 'Asymmetric Paternalism'." University of Pennsylvania Law Review 151 (3): 101-144.

Camerer, Colin, and Richard H. Thaler. 1995. "Anomalies: Ultimatums, Dictators, and Manners." *The Journal of Economic Perspectives* 9 (2): 209-219.

Card, David, Stefano DellaVigna, and Ulrike Malmendier. 2011. "The role of theory in field experiments." *Journal of Economic Perspectives* 25 (3): 39-62.

Carlsson, Fredrik, Peter Frykblom, and Carl Johan Lagerkvist. 2005. "Using cheap talk as a test of validity in choice experiments." *Economics Letters* 89 (2): 147-152.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren. 1990. "A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places." *Journal of Personality and Social Psychology* 58 (6): 1015-1026.

Convery, Frank J. 2009. "Origins and development of the EU ETS." *Environmental and Resource Economics* 43: 391-412.

Costa, Dora L. and Matthew E. Kahn. 2013. "Energy conservation 'nudges' and environmentalist ideology: Evidence from a randomized residential electricity field experiment." *Journal of the European Economic Association* 11 (3): 680-702.

Crocker, Thomas D., Jason F. Shogren, and Paul R. Turner. 1998. "Incomplete beliefs and nonmarket valuation." *Resource and Energy Economics* 20 (2): 139-162.

Crump, Larry S., Dennis L. Nunes, and E. K. Crossman. 1977. "The effects of litter on littering behavior in a forest environment." *Environment and Behavior* 9 (1): 137-146.

Cummings, Ronald G., Glenn W. Harrison, and Laura Osborne. 1995. "Can the bias of contingent valuation be reduced? Evidence from the laboratory." Economics Working Paper B-95-03, Dividing of Research, College of Business Administration, University of South Carolina.

Cummings, Ronald G., and Laura O. Taylor. 1998. "Does realism matter in contingent valuation surveys?" *Land Economics* 74 (2): 203-215.

Cummings, Ronald G., and Laura O. Taylor. 1999. "Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method." *American Economic Review* 89 (3): 649-665.

Dannenberg, Astrid, Andreas Löschel, Gabriele Paolacci, Christiane Reif, and Alessandro Tavoni. 2014. "On the provision of public goods with probabilistic and ambiguous thresholds." *Environmental and Resource Economics* 61: 365-383.

Dawnay, Emma, and Hetan Shaw. 2011. *Behavioural economics: Seven principles for policy makers.* London: New Economics Foundation <http://www.neweconomics.org/publications/entry/behavioural-economics>.

Deaton, Angus. 2010. "Instruments, randomization, and learning about development." *Journal of Economic Literature* 48: 424-455.

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

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for altruism and social pressure in charitable giving." *Quarterly Journal of Economics* 127 (1): 1-56.

Desvousges, William H., F. Reed Johnson, Richard W. Dunford, Kevin J. Boyle, Sara P. Hudson, and K. Nicole Wilson. 1992. *Measuring nonuse damages using contingent valuation: An experimental evaluation of accuracy.* Research Triangle Park, NC: Research Triangle Institute.

Dolan, Paul, Michael Hallsworth, David Halpern, David King, Robert Metcalfe, and Ivo Vlaev. 2012. "Influencing behavior: The mindspace way." *Journal of Economic Psychology* 33 (1): 264-277.

Dolan, Paul, and Robert Metcalfe. Mimeo. "Neighbors, knowledge, and nuggets: Two natural field experiments on the role of incentives on changing energy consumption."

Duflo, Esther. 2006. "Field experiments in development economics." *Advances in Economics and Econometrics: Theory and Applications*, Ninth World Congress, Cambridge University Press.

Dur, Robert, and Ben Vollaard. 2012. "The power of a bad example: A field experiment in household garbage disposal." *Tinbergen Institute Discussion Paper*.

Enkvist, Per-Anders, Tomas Nauclér, and Jerker Rosander. 2007. "A cost curve for greenhouse gas reduction." *McKinsey Quarterly* 1: 34-45.

Farber, Henry S. 2008. "Reference-dependent preferences and labor supply: The case of New York City taxi drivers." *American Economic Review* 98 (3): 1069-1082.

Fehr, Ernst, and Simon Gächter. 2000. "Fairness and retaliation: The economics of reciprocity." *The Journal of Economic* Perspectives 14 (3): 159-181.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl. 1993. "Does fairness prevent market clearing? An experimental investigation." *Quarterly Journal of Economics* 108 (2): 437-459.

Ferraro, Paul, Juan Jose Miranda, and Michael Price. 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.

Ferraro, Paul, and Michael Price. 2013. "Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment." *Review of Economics and Statistics* 95 (1): 64-73.

Geller, Scott E., Jill F. Witmer, and Margaret A. Tuso. 1977. "Environmental interventions for litter control." *Journal of Applied Psychology* 62 (3): 344-351.

Gintis, Herbert. 2000. "Beyond *Homo economicus*: Evidence from experimental economics." *Ecological Economics* 35 (3): 311-322.

Glaeser, Edward L. 2006. "Paternalism and psychology." *University of Chicago Law Review* 73 (1): 32-38.

Glewwe, Paul, and Michael Kremer. 2006. "Schools, teachers, and education outcomes in developing countries." *Handbook on the Economics of Education* 2, pp. 945-1018.

Gneezy, Uri, and John A. List. 2006. "Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments." *Econometrica* 74 (5): 1365-1384.

Gong, Min, and David Aadland. 2011. "Interview effects in an environmental valuation telephone survey." *Environmental and Resource Economics* 49 (1): 47-64.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. "An experimental analysis of ultimatum bargaining." *Journal of Economic Behavior and Organization* 3 (4): 367-388.

Hardin, Garrett. 1968. "The tragedy of the commons." *Science* 162 (3859): 1243-1248.

Harrison, Glenn W., and John A. List. 2004. "Field experiments." *Journal of Economic Literature* 42 (4): 1009-1055.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the case for social experiments." *The Journal of Economic Perspectives* 9 (2): 85-110.

Hepburn, Cameron, Stephen Duncan, and Antonis Papachristodoulou. 2010. "Behavioural economics, hyperbolic discounting, and environmental policy." *Environmental and Resource Economics* 46 (2): 189-206.

Herberich, David H., John A. List, and Michael K. Price. 2011. "How many economists does it take to change a light bulb? A natural field experiment on technology adoption." Working Paper.

Herriges, Joseph, Catherine Kling, Chih-Chen Liu, and Justin Tobias. 2010. "What are the consequences of consequentiality?" *Journal of Environmental Economics and Management* 59 (1): 67-81.

Holland, Rob W., Henk Aarts, and Daan Langendam. 2006. "Breaking and creating habits on the working floor: A field-experiment on the power of implementation intentions." *Journal of Experimental Social Psychology* 42 (6): 776-783.

Hopper, Joseph, and Joyce Nielsen. 1991. "Recycling as altruistic behavior: Normative and behavioral strategies to expand participation in a community recycling program." *Environment and Behavior* 23 (2): 195-220.

Horowitz, John K., and Kenneth E. McConnell. 2002. "Willingness to accept, willingness to pay, and the income effect." *Journal of Economic Behavior and Organization* 51 (4): 537-545.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. 2015. "The persistence of moral suasion and economic incentives: Field experimental evidence from energy demand." *National Bureau of Economic Research* Working Paper No. w20910.

Jaffe, Adam B., and Robert N. Stavins. 1995. "Dynamic incentives of environmental regulations: The effects of alternative policy instruments on technology diffusion." *Journal of Environmental Economics and Management* 29: S43-S63.

Jessoe, Katrina, and David Rapson. 2012. "Knowledge is (less) power: Experimental evidence from residential energy use." *American Economic Review* 104 (4): 1417-1438.

Johnson, Eric J., and Daniel Goldstein. 2003. "Do Defaults Save Lives?" *Science* 302 (5649): 1338-1339.

Kahn, Matthew E., and Frank A. Wolak. 2013. "Using information to improve the effectiveness of nonlinear pricing: Evidence from a field experiment." Working paper.

Kahneman, Daniel, and Jack L. Knetsch. 1992. "Valuing public goods: The purchase of moral satisfaction." *Journal of Environmental Economics and Management* 22 (1): 57-70.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1990. "Experimental tests of the endowment effect and the Coase Theorem." *Journal of Political Economy* 98 (6): 1325-1348.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1991. "Anomalies: The endowment effect, loss aversion, and status quo bias." *Journal of Economic Perspectives* 5 (1): 193-206.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect theory: An analysis of decision under risk." *Econometrica* 47 (2): 263-292.

Kahneman, Daniel, and Amos Tversky. 1981. "The framing of decisions and the psychology of choice." *Science* 211 (4481): 453-458.

Karlan, Dean, and John A. List. 2007. "Does price matter in charitable giving? Evidence from a large-scale natural field experiment." *American Economic Review* 97 (5): 1774-1793.

Keizer, Kees, Siegwart Lindenberg, and Linda Steg. 2008. "The spreading of disorder." *Science* 322 (5908): 1681-1685.

Kesternich, Martin, Andreas Löschel, and Daniel Römer. 2014. "The long-term impact of matching and rebate subsidies when public goods are impure: Field experimental evidence from the carbon offsetting market." *CAWM Discussion Paper*, Centrum für Angewandte Wirtschaftsforschung Münster, no. 76.

Kolstad, Charles. 2010. "Equity, heterogeneity and international environmental agreements." *The B.E. Journal of Economic Analysis and Policy* 10 (2).

Krauss, Robert M., Jonathan L. Freedman, and Morris Whitcup. 1978. "Field and laboratory studies of littering." *Journal of Experimental Social Psychology* 14: 109-122.

Kremer, Michael. 2003. "Randomized evaluations of educational programs in developing countries: Some lessons." *American Economic Review* 93 (2): 102-106.

Krendl, Kathya, Beth Olson, and Richard Burke. 1992. "Preparing for the environmental decade: A field experiment on recycling behavior." *Journal of Applied Communication Research* 20 (1): 19-36.

Kunreuther, Howard, and Elke U. Weber. 2014. "Aiding decision making to reduce the impacts of climate change." *Journal of Consumer Policy* 37: 397-411.

Laibson, David. 1997. "Golden eggs and hyperbolic discounting." *Quarterly Journal of Economics* 112 (2): 443-477.

Landry, Craig E., and John A. List. 2007. "Using ex ante approaches to obtain credible signals for value in contingent markets: Evidence from the field." *American Journal of Agricultural Economics* 89 (2): 420-429.

Ledyard, John O. 1995. "Public goods: A survey of experimental research." In *Handbook of Experimental Economics*, eds. John Kagel and Alvin E. Roth. Princeton: Princeton University Press, 111-194.

Leggett, Christopher G., Naomi S. Kleckner, Kevin J. Boyle, John W. Duffield, and Robert Cameron Mitchell. 2003. "Social desirability bias in contingent valuation surveys administered through in-person interviews." *Land Economics* 79 (4): 561-575.

Liebig, Georg, and Jens Rommel. 2014. "Active and forced choice in overcoming status quo bias: A field experiment on the adoption of "No junk mail" stickers in Berlin, Germany." *Journal of Consumer Policy* 37: 423-435.

List, John A. 2001. "Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards." *American Economic Review* 91 (5): 1498-1507.

List, John A. 2003. "Does market experience eliminate market anomalies?" *Quarterly Journal of Economics* 118: 41-71.

List, John A. 2011a. "Why economists should conduct experiments and 14 tips for pulling one off." *Journal of Economic Perspectives* 25 (3): 3-16.

List, John A. 2011b. "Does market experience eliminate market anomalies? The case of exogenous market experience." *American Economic Review* 101 (3): 313-317.

List, John A., Robert P. Berrens, Alok K. Bohara, and Joe Kerkvliet. 2004. "Examining the role of social isolation on stated preferences." *American Economic Review* 94 (3): 741-752.

List, John A., and David Lucking-Reiley. 2002. "The effects of seed money and refunds on charitable giving: Experimental evidence from a university capital campaign." *The Journal of Political Economy* 110 (1): 215-233.

List, John A., and Michael K. Price. 2013. "Using Field Experiments in Environmental and Resource Economics." *National Bureau of Economic Research* Working Paper No. 19289.

List, John A., and Daniel M. Sturm. 2006. "How elections matter: Theory and evidence from environmental policy." *Quarterly Journal of Economics* 121 (4): 1249-1281.

Löfgren, Asa, Peter Martinsson, Magnus Hennlock, and Thomas Sterner. 2012. "Are experienced people affected by a pre-set default option? Results from a field experiment." *Journal of Environmental Economics and Management* 63: 66-72.

Loureiro, Maria L., and Justus Lotade. 2005. "Do fair trade and eco-labels in coffee wake up the consumer conscience?" *Ecological Economics* 53: 129-138.

Lusk, Jayson L. 2003. "Effects of cheap talk on consumer willingness-to-pay for golden rice." *American Journal of Agricultural Economics* 85 (4): 840-856.

McCaul, Kevin D., and John T. Kopp. 1982. "Effects of goal setting and commitment on increasing metal recycling." *Journal of Applied Psychology* 67 (3): 377-379.

Meier, Stephan. 2007. "Do subsidies increase charitable giving in the long run? Matching donations in a field experiment." *Journal of the European Economic Association* 5 (6): 1203-1222.

Milinski, Manfred, Ralf D. Sommerfeld, Hans-Jürgen Krambeck, Floyd A. Reed, and Jochem Marotzke. 2008. "The collective-risk social dilemma and the prevention of simulated dangerous climate change." *Proceedings of the National Academy of Sciences* 105 (7): 2291-2294.

Mullainathan, Sendhil, and Richard H. Thaler. 2000. "Behavioral economics." *National Bureau of Economic Research* Working Paper No. 7948.

Nordhaus, William D. 2007. "A review of the *Stern Review on the Economics of Climate Change*." *Journal of Economic Literature* 45 (3): 686-702.

Ramos, Joao, and Benno Torgler. 2012. "Are academics messy? Testing the broken windows theory with a field experiment in the work environment." *Review of Law and Economics* 8 (3): 563-577.

Reisch, Lucia, and Cass R. Sunstein. 2014. "Redesigning cockpits: Introduction to special issue of Journal of Consumer Policy on behavioural economics, environmental policy and the consumer." *Journal of Consumer Policy* 37: 333-339.

Reiter, Susan M., and William Samuel. 1980. "Littering as a function of prior litter and the presence or absence of prohibitive signs." *Journal of Applied Social Psychology* 10 (1): 45-55.

Schultz, P. Wesley. 1999. "Changing behavior with normative feedback interventions: A field experiment on curbside recycling." *Basic and Applied Social Psychology* 21 (1): 25-36.

Shogren, Jason. 2012. "Behavioural economics and environmental incentives." *OECD Environment Working Papers*. No. 49. OECD Publishing.

Shogren, Jason F., and Laura O. Taylor. 2008. "On behavioral-environmental economics." *Review of Environmental Economics and Policy* 2 (1): 26-44.

Smith, Vernon L. 1982. "Microeconomic systems as an experimental science." *American Economic Review* 72 (5): 923-955.

Stavins, Robert N. 1998. "Economic incentives for environmental regulation." In *The New Palgrave Dictionary of Economics and the Law*, ed. Peter Newman. London: Macmillan Reference.

Stern, Nicholas. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge University Press.

Stoop, Jan, Charles N. Noussair, and Daan van Soest. 2012. "From the lab to the field: Cooperation among fishermen." *Journal of Political Economy* 120 (6): 1027-1056.

Sunstein, Cass R. 2011. "Empirically informed regulation." *The University of Chicago Law Review* 78 (4): 1349-1429.

Tavoni, Alessandro, Astrid Dannenberg, Giorgos Kallis, and Andreas Löschel. 2011. "Inequality, communication, and the avoidance of disastrous climate change in a public goods game." *Proceedings of the National Academy of Sciences* 108 (29): 11825-11829.

Tavoni, Alessandro, and Simon Levin. 2014. "Managing the climate commons at the nexus of ecology, behaviour, and economics." *Nature Climate Change* 4 (12): 1057-1063.

Thaler, Richard. 1999. "Mental accounting matters." *Journal of Behavioral Decision Making* 12: 183-206.

Thaler, Richard, and Cass Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press: New Haven, CT.

Tiefenbeck, Verena, Thorsten Staake, Kurt Roth, and Olga Sachs. 2013. "For better or worse? Empirical evidence of moral licensing in a behavioral energy conservation campaign." *Energy Policy* 57: 160-171.

Tversky, Amos, and Daniel Kahneman. 1973. "Availability: A heuristic for judging frequency and probability." *Cognitive Psychology* 5 (2): 207-23.

Vossler, Christian A., Maurice Doyon, and Daniel Rondeau. 2012. "Truth in consequentiality: Theory and field evidence on discrete choice experiments." *American Economic Journal: Microeconomics* 4 (4): 145-171.

Wolak, Frank A. 2006. "Residential customer response to real-time pricing: The Anaheim critical peak pricing experiment." Working Paper, *Center for the Study of Energy Markets*.

Wolak, Frank A. 2011. "Do residential customers respond to hourly prices? Evidence from a dynamic pricing experiment." *American Economic Review* 101 (3): 83-87.

Yoeli, Erez, Moshe Hoffman, David G. Rand, and Martin A. Nowak. 2013. "Powering up with indirect reciprocity in a large-scale field experiment." *Proceedings of the National Academy of Sciences* 110 (2): 10424-10429.

CHAPTER II

BE WHO YOU *OUGHT* OR BE WHO YOU *ARE*?
ENVIRONMENTAL FRAMING AND COGNITIVE DISSONANCE IN GOING PAPERLESS

By Greer Gosnell

Abstract: We manipulate message framing to analyze behavioral motivators businesses may consider when encouraging customers—here, those with revealed environmental preferences—from paper billing to online billing. In a large-scale natural field experiment comprising 38,654 Good Energy customers in total, we investigate a role for targeted messaging based on consumer preferences and beliefs in emails promoting an active switch to paperless billing. Through randomization of environmental information and messaging rooted in theories of cognitive dissonance—a phenomenon centered upon a desire for consistency in self-perception—we find that environmental information is ineffective in inducing behavior change. Interestingly, the cognitive dissonance treatment weakly improves uptake among our main sample but largely backfires among a subsample of individuals with extensive postgraduate education. Contrary to the majority of the literature on gender and environmental behavior, females are less likely to switch to paperless billing.

Acknowledgments: Thank you to Juliet Davenport for her passionate interest the use of field experiments to further the mission of Good Energy. Thanks also to her team—especially Dave Ford and Will Vooght—for their valuable time and effort in the design and implementation of this study.

1. Introduction

Businesses and governments are increasingly turning to randomized experiments to discover means by which to increase profitability or pursue policy goals. In a number of contexts, such social and private objectives coincide, creating opportunity for partnerships between academic researchers and businesses interested in either or both of said objectives. For instance, Opower—a thriving for-profit energy information provider—was founded upon robust research originally intended to help energy suppliers transition their business models to increase customer satisfaction and retention; the customer-centric approach simultaneously helps the utilities' customers to scale back on inefficient energy consumption in the home, thereby reducing environmentally costly greenhouse gas emissions. Over the last decade, Opower has worked alongside utilities and academic researchers to implement randomized experiments that demonstrate the effectiveness of their product, i.e. provision of tailored social norm information to householders (e.g., see Allcott, 2011).

Indeed, customer communications is of paramount importance to businesses in terms of both allocation of costly resources as well as customer retention and satisfaction. As a means of increasing the efficiency of operations, the business world has seen a clear and rapid capitalization upon technological advancements—such as mobile phone applications and text messaging, or automatic bill pay (ABP)—in endeavors to simplify and expedite everyday business practices. However, enrollment in such programs may lead to welfare loss, as demonstrated in Sexton (2015). In this paper, residential customers enrolled in ABP increase energy consumption by 4.0% on average, an effect that increases to 7.3% for small- to medium-sized commercial and municipal customers. Thus, while enrolling customers in alternative bill payment schemes may decrease transaction costs for retailers and improve resourcefulness, the act may come at a cost in terms of customer satisfaction and ultimately retention. Instead, companies may offer the option to switch, but status quo bias and potential costs (e.g., from increased consumption, as in Sexton, 2015) suggest that many consumers may refrain from opting in.

How can companies maximize opt-in rates for resourceful changes in communications? In this paper, we investigate means by which to facilitate such cost- and resource-efficient change without forcing the change upon the customer. We manipulate message framing to analyze behavioral motivators businesses may consider when encouraging customers—here, those with

revealed environmental preferences—from paper billing to online billing. We investigate a role for targeted messaging based on consumer preferences and beliefs through randomization of environmental information and messaging rooted in theories of cognitive dissonance—a phenomenon centered upon a desire for consistency in self-perception. Our design relies on the assumption that the customer base of Good Energy, a 100% renewable energy supplier in the United Kingdom, is characterized by strong environmental preferences. The assumption is founded upon the mission of Good Energy "to keep the world a habitable place by offering consumers an active role in addressing climate change."[7] In light of the social mission of Good Energy and its customers' selection into their customer base, we conceptualize a utility function susceptible to information and cognitive dissonance, designing interventions to manipulate arguments in the utility function related to social preferences and self-perception.

Neoclassical economics holds that information influences behavior through its effects on individuals' internal cost-benefit analyses, which are rooted in preferences characterized by selfishness (DellaVigna, 2009). More recent economic theories posit that such cost-benefit analyses incorporate altruistic preferences (Becker, 1974; Andreoni, 1989, 1990). Theories in social psychology draw similar conclusions regarding the role of information on attitudes and behaviors (e.g., Ajzen and Fishbein, 1980; Stern, 2000; Kollmuss and Agyeman, 2002). Perhaps counter to intuition, there is ample empirical evidence that calls into question the effectiveness of information in changing human behavior. We test whether social information influences the decision making of individuals with revealed environmental preferences. That is, our first intervention aims to promote paperless billing through provision of information on environmental costs associated with paper use.

More recently, following the publication of the seminal research by Kahneman and Tversky (1979), psychological 'irregularities' have begun to permeate economic theory; economists have since incorporated empirical patterns such as loss aversion, anchoring, and probability weighting into traditional utility functions. As evidenced in Section II below, cognitive dissonance is no exception. While several cognitive dissonance theories have been proposed and (to some extent) tested, field experimentalists have arguably underutilized the psychological phenomenon, whether as a means to explain behaviors inconsistent with neoclassical economic predictions or as a vehicle for behavior change. Our second intervention

[7] http://www.goodenergygroup.co.uk/about/mission

investigates the role that innate desire for consistency across one's beliefs and behaviors may play in encouraging conservation.

Finally, a sparse literature appears to suggest that imagery can induce behavior change. For instance, a series of lab experiments (Haley and Fessler, 2005; Burnham and Hare, 2007; Rigdon et al., 2009; Mifune, Hashimoto, and Yamagishi, 2010) and field experiments on honesty, littering, and donating (Bateson, Nettle, and Roberts, 2006; Ernest-Jones, Nettle, and Bateson, 2011; Ekström, 2012) demonstrate that an image of eyes can cause individuals to comply with cooperative norms in some contexts. Additionally, money priming has been shown to lead people to make less altruistic decisions or to focus their attention on monetary features of products (see Vohs, 2015, for a review). Here, we test whether images of the environmental good under threat (i.e. trees) can serve as a visual reminder of the externality associated with subjects' inaction and therefore increase their probability of acting.

In a large-scale natural field experiment comprising 38,654 Good Energy customers, we randomize environmental information, dissonance-inducing messaging, and environmental imagery in emails promoting an active switch to paperless billing.[8] In addition to household-level data on e-billing sign-up, our data include both gender and a proxy for level of education (i.e. whether the householder is a 'Doctor' or 'Professor'), two demographic factors that have been shown to increase pro-environmental behavior (Kollmuss and Agyeman, 2002). We find that both imagery and information on environmental costs associated with the status quo are ineffective in increasing uptake of paperless billing beyond that of a control group. On the other hand, dissonance-inducing messaging increases uptake among our main sample; interestingly, it backfires among our highly educated sample. We attribute the latter finding to a lesser need for reassurance of a moral (e.g., socially-minded) self-concept.[9] Additionally, we find that women are less likely than men to sign up to paperless billing. The research suggests that individuals may be targeted with various forms of messaging to increase low-cost environmentally advantageous behaviors, and calls into question the general conclusion in the literature that women are more inclined than men to behave in line with environmental objectives.

This paper is structured as follows. Section 2 provides background on the mechanisms investigated in our treatments, namely the role of information in environmental decision making and the infusion of cognitive dissonance into the study of economic decision making. Section 3

[8] Emails were sent to all customers for whom Good Energy had an email address on file (i.e. 84% of accounts).
[9] Potential alternative explanations include a relatively strong priority for convenience, or knowledge of (and subsequent non-susceptibility or aversion to) marketing tactics rooted in psychological phenomena.

outlines the experimental design and details the interventions implemented across Good Energy's customer base. Section 4 reveals the results of the field experiment, and Section 5 concludes.

2. Background and motivation

i. Be who you ought: Information provision

While the rational economic man of neoclassical theory is influenced by two primary motivators—information and incentives—social psychology and behavioral economics reserve a role for evaluative, normative, and identity-driven beliefs and motivations (Ajzen and Fishbein, 1980; Elster, 1989; Akerlof and Kranton, 2000; Stern, 2000; Kollmuss and Agyeman, 2002). According to the norm-activation theory of Schwartz (1973) and the value-belief-norm (VBN) theory of Stern et al. (1999), knowledge of negative consequences associated with one's actions—or particular undesirable conditions for which one is perceived to be responsible—spurs altruistic behavior. Therefore, information regarding particular externalities (or internalities) may change individuals' beliefs and intentions, in turn altering their proclivity to engage in socially or personally beneficial behaviors (Stern, 2000).[10]

Empirically speaking, and despite the overwhelming tendency of social campaigns to communicate information with the goal of changing behavior, the impact of consequence-based information on subsequent behavior has proven negligible in a number of settings.[11] In a Norwegian experiment, parents were randomly assigned to receive a short informational briefing and brochures on smoking and its harmful passive effects on their children during well-child appointments, and self-reported smoking behavior did not change (Eriksen, Serrum, and Bruusgaard, 1996).[12] Similarly, several studies demonstrate a non-effect of information—including calories per item and recommended daily caloric intake—on subsequent order choice in fast food restaurants (Harnack et al., 2008; Downs, Loewenstein, and Wisdom, 2009). Likewise, extreme media coverage of the consequences of Enron's accounting scandal on 401(k) holdings

[10] While Stern et al. (1999) find empirical support for their VBN theory, Kollmuss and Agyeman (2002) argue that the determinants of pro-environmental behavior are so varied and complex—dependent on myriad interactions between internal and external influences, as well as demographic factors[10]—that no model could possibly predict its (non-) occurrence. Furthermore, they claim that the effect of environmental knowledge and awareness is highly contextual and rests on the preferences and beliefs of the individual to which information is imparted, and the ability of that information to trigger emotional involvement.

[11] Zelezny (1999) finds that educational interventions—particularly in classrooms—can improve environmental behaviors, but the meta-analysis comprises a number of studies with poor methodology.

[12] In a Dutch experiment, information on the personal and social consequences of smoking altered perceptions of the experience of smoking in a negative direction (Dijkstra et al., 1998). Subjects construed the experience as less tasteful and pleasurable, though any resulting changes in behavior were not measured.

did not prompt employees in similar companies to diversify their 401(k) investments (Choi, Laibson, and Madrian, 2005).[13]

The consequences discussed above are primarily 'internalities', or unintended costs of one's actions that accrue to oneself alone. A meta-analysis of interventions intended to reduce household energy consumption demonstrates that information regarding externalities may increase knowledge but does not subsequently alter behavior (Abrahamse et al., 2005). On the other hand, Ferraro and Price (2011) find that information on the extent and consequences of water use among its (environmentally unconscious) customer base increased the implementation of water-saving strategies, especially among high-consumption households. Additionally, using a randomized field experiment in Brazilian favelas, Toledo (2016) finds that environmental persuasion increases take-up of LED (energy-efficient) light bulbs by 6 percentage points (or 13%). In contrast to our setting, the outcome of interest in these cases is costly, as they require that individuals actively change their habits or spend money to reduce their energy and water consumption.

In addition, the interventions are applied to individuals who do not necessarily exhibit a preference for the healthy or financially advantageous outcomes that constitute the focus of those studies. In a Dutch mass media campaign surrounding the causes of and possible behavioral solutions for climate change, individuals who reported a higher willingness to engage in pro-environmental behaviors were those who had already been behaving in such a manner prior to the campaign (Staats, Wit, and Midden, 1996). That is, information campaigns may be more effective in inducing behavior change among individuals already motivated prior to intervention.

We explore a role for information regarding environmental externalities on a targeted audience of individuals exhibiting green preferences, where the information provided is directly and specifically related to the outcome behavior of interest. As environmental issues become more prominent in media and education, this environmentally conscious audience is growing and is arguably the segment of the population most inclined to change their behavior as a result of exposure to information on environmental externalities for which they are (partially) responsible; such individuals tend to possess a locus of control and have likely already acted prosocially in

[13] Several additional experiments demonstrate the negligible effects of financial education. For instance, in a randomized experiment, information provided at a benefits information fair was shown to have only a small (though significant) positive effect on employee enrollment in a Tax Deferred Account scheme in a university setting (Duflo and Saez, 2003). Similarly, hour-long financial education seminars at a large insurance provider proved to stimulate only marginal improvements in 401(k) investment (Madrian and Shea, 2001; Choi, Laibson, Madrian, and Metrick, 2002).

accordance with their environmental knowledge in signing up to this particular utility. Unlike many studies on the effects of information, our setting controls for any external influences (e.g., economic or social incentives) and targets an extremely low-cost behavior—namely making a one-time switch from paper billing to online billing—so that attitudinal factors likely play a causal role in decision making (Stern, 2000).

ii. Be who you are: Cognitive dissonance

Theories of cognitive dissonance originated in psychology and have since piqued the interest of a number of economists. The theories generally rest upon the assumption that human beings are averse to inconsistencies between past or current beliefs and behaviors (Festinger, 1957). In general, individuals strive for consistency, competence, and morality in their perceptions of themselves, and behaving in a manner that negates these features results in psychological discomfort (Aronson, 1992). Such 'dissonance' is morally costly, and economic agents will incorporate these costs into their utility maximization problems (e.g., Gilad, Kaish, and Loeb, 1984; Konow, 2000). Hence, cognitive dissonance may be responsible for some portion of the 'noise' inherent in traditional neoclassical models.

According to Gilad, Kaish, and Loeb (1987), cognitive dissonance can manifest in situations in which "a decision is undertaken freely and with the understanding of possible adverse outcomes" (p. 64). In their theory of selective exposure, behavior remains consistent with traditional utility maximization if exposure to certain types of information can be controlled and dissonance kept at a level below some threshold. Otherwise, the individual must change her beliefs (which is costly), and she will subsequently maximize in accordance with a revised objective function.

Rabin (1994) proposes a similar structure for the utility function, adding a more nuanced explanation of the contexts in which cognitive dissonance will increase the tension between material benefit and psychological cost. For instance, he conjectures that an individual who receives less material benefit from an immoral activity will further convince himself of the immorality of the activity. Interestingly, he shows that a stronger proclivity toward cognitive dissonance may pressure an individual with high material benefit from said activity into changing her beliefs, thereby augmenting immoral activity. In his primary example, he discusses the acceptance of ethically questionable activity in the meat industry, asserting that strong moral

indoctrination combined with high utility from meat consumption will lead to internal justification of otherwise unethical practices (i.e. a change in beliefs). Counter to intuition, the agglomerated effects of this phenomenon may mean that inculcating members of society with strong moral beliefs could eventually lead to higher levels of immoral activity.

A more recent interpretation of cognitive dissonance also emphasizes the role of context in determining the extent to which one may rationalize decisions in light of her beliefs. Mazar, Amir, and Ariely (2008) put forth a theory of cognitive dissonance in which the propensity to engage in dishonest behavior is dependent on two types of contextual feature. The first—'categorization'—refers to the extent to which the given context facilitates flexibility of interpretation with respect to self-perception, or the extent to which the act may plausibly be considered consistent with the self-concept (which may, in turn, depend on the strength and relevance of social norms; see Wichardt, 2012). For instance, Nail, Misak, and Davis (2004) point out that the dissonance-inducing act must be voluntary and otherwise unjustifiable (or difficult to justify), and must involve perception of commitment. The second type relates to individuals' mindfulness of and attention to their own moral standards. In several laboratory experiments, they find that individuals who have the opportunity to cheat do so, though they are less likely to cheat when reminded of their moral beliefs or after signing an honor code. They argue that the salience of self-concept is, therefore, an important driver of congruence between belief and behavior.

In a more formal economic theory, Akerlof and Dickens (1982) propose a two-period model in which a rational individual first chooses whether to participate in a safe or hazardous industry; if she chooses the latter, she will convince herself of the safety of the industry so as to justify her past decision. In the second period, a cost-effective safety device becomes available and the individual—who would have purchased the device had it been available prior to her perception change—continues to work without it. According to the authors, their model justifies government intervention requiring hazardous industry workers to wear the equipment in order to return to Pareto optimal conditions. While the model focuses on labor selection, it is also applied to explain the effectiveness of non-informational advertising, the incidence of crime under various degrees of sanctions, and the necessity of Social Security for individuals who are averse to acknowledging the inevitability of old age.

Finally, Konow (2000) posits a utility function comprising material wealth along with two costly parameters: cognitive dissonance and self-deception. The former characterizes the deviation between one's beliefs and one's actions—in this case, the deviation between a fair allocation and one's actual allocation in a dictator game—while the latter captures the discomfort associated with altering one's initial fairness perspective to increase consistency between the aforementioned allocations. Experimental results from several variants of the dictator game—where subjects perform both active and passive dictator roles—provide strong empirical support for both phenomena.

Do individuals express opinions or take part in costly activities in order to remain consistent with self-perceptions outside of the laboratory? Can cognitive dissonance explain sacrifice for the sake of fairness in the real world? Indeed, social scientists have cited cognitive dissonance as an explanation for voting behavior (Mullainathan and Washington, 2006), investor inertia (Goetzmann and Peles, 1997; Rennekamp, Rupar, and Seybert, 2015), sexual risk taking (Mannberg, 2012), diminished labor supply in the face of job search discrimination (Goldsmith et al., 2004), endogenous class formation (Oxoby, 2003), and honesty in the face of cheating opportunities (Mazar, Amir, and Ariely, 2008). Furthermore, the phenomenon has been exploited as a means to ends such as water use reduction (Dickerson et al., 1992), sustained weight loss (Axsom and Cooper, 1985), and reducing hypothetical bias in contingent valuation studies (Alfnes, Yue, and Jensen, 2010).

We extend the above literatures in a somewhat new direction. Using a large-scale natural field experiment, we test a role for information provision and cognitive dissonance in encouraging renewable gas and energy consumers to switch from resource-intensive paper billing to online billing. To investigate a role for information provision in influencing resourceful behavior, we provide the utility customers with information on the social consequences of continuing to receive communications in the post. To test the impact of cognitive dissonance on e-billing take-up, we promote present decision making consistent with implicit beliefs associated with related past decisions. In sum, we implement treatments that both appeal to embedded environmental preferences and that target preferences for a consistent self-perception.

In light of the above theories, we hypothesize that 'green' consumers of a renewable energy utility will respond to the cognitive dissonance intervention by switching from paper billing to online billing if the cost of such dissonance sufficiently outweighs the convenience to

the consumer of paper billing and any perceived cost savings associated with its salience. Additionally, in line with VBN theory and theories of identity, we posit that information on environmental damage will trigger motivation to act altruistically, especially if individuals have internalized the norms of eco-consciousness associated with being a Good Energy customer. Finally, in line with Taylor and Thompson (1982)'s conclusion that vividness may be important in the context of everyday informational competition, we conjecture that environmental imagery may serve to enhance the salience of environmental costs, thereby augmenting the perceived benefits of taking action and increasing the probability of doing so.

3. Experimental design

We partnered with Good Energy—the UK's leading renewable energy supplier—to randomize email content in a campaign to encourage customers to switch from their current information channel (i.e. quarterly paper bills received by mail) to online billing (i.e. quarterly bills received via email). The six-week campaign ran in September and October of 2014.

As a business founded upon an environmental mission, Good Energy's objective was to achieve a switch rate as close to 100% as possible. Additionally, online billing constitutes a cost reduction, as online billing requires less physical and human capital than does paper billing. We test the effectiveness of information on environmental savings ('environmental framing') and a reminder of the customers' environmental preferences ('cognitive dissonance'). Each email begins by announcing the arrival of the e-billing option and is sent from Good Energy's Chief Operating Officer. The defining features of each email intervention are detailed below.

i. Interventions

Control (Groups 1-2). In the control email, the first line unveils the online billing option ('availability line' hereafter, emphasis included): "**It's finally here!** Now you can switch to e-billing and have your energy bills emailed directly to your inbox rather than receiving them by post." The subsequent line touts online billing access ('online access line' hereafter): "Even better, you can access your bills online any time, so they won't fill any valuable space in your drawers or bins." Both of the previous lines appear identically across all interventions.

The key following control statement reads, "**Here at Good Energy, we prioritise customer satisfaction.** The opportunity to switch to e-billing is just one more step we have taken to keep you smiling." Then, three benefits of switching are listed: 1) Reduce paper waste; 2) Spend less time sorting through mail; and 3) Access bills 24/7 online. The email includes a link to make the switch, and all emails contain the same closing statement followed by a signature from the Chief Operating Officer (for full letter, see Figure A1 in Appendix).

Environmental Framing (Groups 3-4). This treatment provides information on the environmental benefits associated with a universal shift of GE customers to e-billing. Following the availability line stated above, this treatment states (emphasis included), "**If all customers make the switch, we would save 46 trees worth of paper each year!**" This line is followed by the online access line.

In addition to emphasizing GE's attention to customer satisfaction, the next line also points out its commitment to the environment (emphasis included): "**Here at Good Energy, we prioritise customer satisfaction as well as the environment.** The opportunity to switch to e-billing is just one more step we have taken to keep you smiling and help you shrink your environmental footprint." The subsequent benefits no longer appeal to the customer herself, but rather are informative of the extent of paper waste and its environmental costs. The first bullet states, "The average UK family throws away 6 trees worth of paper in their household bin each year." The second pertains to the energy and climate impacts of the paper industry as a whole: "Paper production ranks 3rd and 4th for most energy intensive and greenhouse gas intensive manufacturing industries (respectively)." Finally, we provide aggregate paper use statistics for the UK: "12.5 million tonnes of paper and cardboard are used annually in the UK, making us the 11th worst paper offender in the world." The email closes as indicated in the control description (for full letter, see Figure A2 in Appendix).

Control and Environmental Framing (Groups 5-6). While the content contained in the above treatment email is roughly the same length and format as the control email, it contains some fundamentally different information. We also test whether provision of the environmental information (presented to Groups 3 and 4) in addition to the control information (provided to Groups 1 and 2) is effective, allowing us to control for the otherwise substantial change in content from one email to the next (see Table 1). All information from both the control and the

environmental framing email is aggregated into one email (for full letter, see Figure A3 in Appendix).

Cognitive Dissonance (Groups 7-8). Our final treatment quite closely emulates the control email with the exception of a single line difference, so that length and format are quite similar.[14] Instead of emphasizing customer satisfaction, this email appeals to one's identity as a conscious decision maker: "**As a Good Energy customer, you are an environmental steward.** By switching to e-billing, you take another important step to eliminate the environmental impact of your energy use." The remainder of the email is identical to the control email (for full letter, see Figure A4 in Appendix).

Environmental Image (Groups 2, 4, 6, and 8). Finally, we tested the effectiveness of imagery—a central and customary component of Good Energy's communications strategy—in capturing customers' attention. For each of the above treatment emails, an additional treatment intervention existed with the same email content but with a vibrant image of trees at the outset (see Figure A5 in Appendix). All other content in the letters remains identical.

ii. Data: Good Energy customers

The main sample consists of 36,810 Good Energy customers, which is the entire customer base omitting those for whom a working email address had not been provided or for whom gender cannot be identified. This sample is 47% female. The average customer has been with Good Energy for 315 days and consumes 6450 kWh in gas and 3435 kWh in electricity on an annual basis. Customers on a dual fuel account (i.e. who have both gas and electricity accounts with Good Energy) comprise 41% of the sample, while those with gas or electricity only constitute 6% and 53%, respectively. A separate analysis is performed for those identified as either 'Doctor' or 'Professor' and are therefore gender neutral, of which there are 1844 customers (approximately 5% of the sample). Of these customers, the average customer duration is 320 days, average annual gas and electricity consumption are 7592 kWh and 3546 kWh (respectively), and 41%, 7%, and 52% are on dual-fuel, gas, and electricity contracts (respectively). The difference in the two samples is significant for annual gas consumption

[14] Since the length and format are very similar, we do not include a treatment that combines the content of the Control and Cognitive Dissonance interventions. Instead, we simply substitute messaging regarding customer satisfaction for messaging invoking identity as an environmental steward, so that *not* signing up to paperless billing may induce cognitive dissonance.

($p<0.01$) and proportion of gas-only customers ($p<0.10$). We control for all of the above observables in the analysis.

More generally, the customers of Good Energy are fairly representative of UK households more broadly in terms of energy consumption and costs. In our data, the average estimated annual energy consumption is 3,668 kWh, while the average UK household in 2014 consumed 4,001 kWh. On the other hand, Good Energy gas customers use slightly more gas (13,827 kWh) than the average British household (12,404 kWh).[15] Additionally, customers in our data likely pay similar prices per kWh. Due to increased competitiveness of renewable energy in the UK market, Good Energy customers pay a competitive price for their energy. On average, while dual fuel customers of the UK's 'Big Six' energy providers paid approximately £1360 per household in 2013, Good Energy households paid £1313 (see Figure 1). Similarly, compared to Ecotricity, one of Good Energy's primary competitors in the UK renewable energy market, Good Energy dual fuel customers paid £55 less per annum. Therefore, cost of energy does not distinguish Good Energy households from other UK households.

iii. Randomization

All observable variables in the dataset were used in the stratified randomization. Specifically, customers were sorted according to their account's fuel type (gas only, electric only, or dual fuel), their estimated annual consumption (partitioned into quartiles), the length of their contract with Good Energy (partitioned into deciles), and the gender of the account holder (male, female, unidentified). First, we sorted customers according to the three fuel types, and within each fuel type we blocked them according to the estimated annual total consumption quartiles, creating twelve blocks. Having sorted the data into these twelve blocks, we then sorted customers in each block according to duration of existing contract with Good Energy, followed by the account holder's gender. If all blocks had contained at least one customer, this would have created $12\times10\times3=360$ blocks in total. However, there are nine blocks (i.e. combinations of the above variables used for stratification) for which no customer in the dataset is representative, so the stratification created 351 blocks in total. Once the data is sorted according to the existing 351

[15] Goodright, Victoria, and Emily Wilkes. "Chapter 3: Domestic Energy Consumption in the UK between 1970 and 2014." in *Energy Consumption in the UK (2015).*

blocks, a number (1-8) is assigned to each account holder to allocate each customer to one of the eight treatments described above.[16]

Since Good Energy's email server is limited in terms of the volume of emails that can be sent in one day, the trial was planned for six weeks. We tested for pre-experimental equivalence across all group pairs on the above variables as well as the day of week on which the email would be sent, as shown in the balance table (see Tables 2a and 2b).

4. Results

i. Main results

In total, 13.42% of customers signed up for e-billing. In almost all cases, the email without the image outperformed that with the image; while the difference is not statistically significant when comparing all treatments without images to all treatments with images (p=0.122), the difference is significant when comparing the cognitive dissonance treatments with and without images (p=0.054). Simple t-tests do not reveal significant differences across treatments with varying information (see Table A1 in Appendix). Below, we investigate treatment effects using logistical regression analysis.

Our intent-to-treat analysis considers a binary response variable, and we therefore report the results of a logit model (in terms of both logistic coefficients and odds ratios). The regression performed is specified as follows:

$$logit_i = \alpha + \beta_j T_{i,j} + \gamma X_i + e_{it}, \qquad e_{it} \sim \mathrm{N}[0,1]$$

where $logit_i$ is the log of the odds of e-billing sign-up for individual i, α and $\beta_{1,\dots,j}$ with $j \in (1, 2, 3)$ are unknown population parameters to be estimated in the model, $T_{i,j}$ represent the j treatments, and X_i is a vector of control variables including consumption, tariff type, and gender. Controlling for all other variables in the regression, being in treatment group T_j multiplies the

[16] We calculate sample sizes using the formulas provided in List, Sadoff, and Wagner (2011). In our power calculations, we use conventional power and significance levels of (1-β)=0.8 and α=0.05, respectively, and assume equal outcome variance across treatments. Allocating our sample across eight recipient groups places 4,831 individuals into each arm. With at least 4,826 recipients per group, our power calculations show that we can detect treatment effects of 0.057, or about a 1.71% treatment effect.

odds of signing up to e-billing by the exponential of the logistic coefficient β_j—i.e. $\exp(\beta_j)$, which is equivalent to the odds ratio—for recipients of treatment j, and similarly a unit increase in X_i multiplies the odds of sign-up by $\exp(\gamma)$.

Receiving the cognitive dissonance message (without image) multiplies the odds that one signs up to e-billing by exp(0.105)=1.107, i.e. increases the odds by 10.7%, controlling for consumption, tariff type, and gender ($p<0.10$). However, including the image appears to distract from the dissonance-inducing messaging, eliminating the effect altogether. While the odds of sign-up also tend to increase for the treatment groups containing environmental information, we do not have sufficient power to detect such an effect with statistical significance.[17] Contrary to findings in the literature regarding environmental behavior and gender (see Cheng, Woon, and Lynes, 2011, for a review), we find that being female decreases the odds of signing up to paperless billing by 26.5%; as shown in Table A2, this result holds if we run the logit without treatment indicators within the control treatment alone (26.5% reduction in the odds of sign-up, $p<0.01$). There are no significant interaction effects between gender and treatment (see Table A3).

Additionally, it appears that those with smaller observed environmental footprints are more likely to sign up to e-billing. For instance, relative to those on dual-fuel renewable tariffs, the odds of signing up among customers on either gas- or electricity-only tariffs are approximately 40% and 43% lower ($p<0.01$). Finally, for every increase of 1000 kWh in estimated annual gas and electricity consumption, the odds of sign-up decrease by 0.004% ($p<0.10$) and 0.014% ($p<0.01$), respectively. If we assume that being a dual-fuel consumer is indicative of higher environmental preferences than being a single-fuel consumer, and that lower consumption is associated with higher environmental preference, these final two results appear to imply that individuals with greater preference for the environment are more likely to sign up for paperless billing. Of course, we do not have data on household size or income, so consumption may also act as a proxy for wealth as opposed to environmental preference.

[17] We also do not detect an effect of including control information (Groups 5 and 6) in addition to environmental information (Groups 3 and 4).

ii. Heterogeneity

While we find evidence that gender is a significant predictor of our response variable, we do not have a measure of education for the individuals in the main sample. Similarly, we do not have gender data for the 1844 individuals identified with the title of either 'Doctor' or 'Professor'. Therefore, we run our regressions for the two samples independently. In the absence of an all-inclusive continuous or categorical measure for education, we run the same regression as in Table 3 for just the 'postgraduate education' sample, excluding the gender indicator (see Table 4). Contrary to the main sample, the cognitive dissonance intervention quite drastically backfires when we consider doctors and professors only, decreasing the odds of sign-up by 43.0%. Again, provision of statistics on associated environmental damage does not significantly affect the odds of paperless take-up. Consumption does not predict behavior among this subsample, while again being a dual-fuel customer improves the probability that the individual will sign up quite substantially ($p<0.01$).

If we instead run a logistic regression on the full sample that includes interaction terms between assigned treatment and a dummy indicating whether the individual is in the postgraduate education sample, we find a similar result (Table A4). On average, having extensive postgraduate education increases the odds of signing up to e-billing by 32% ($p=0.141$). Without controlling for gender, the odds of signing up to e-billing in the cognitive dissonance (without image) treatment *increase* by 10.7% ($p=0.096$) in the main sample, while the odds *decrease* by 48.7% ($p=0.023$) for doctors and professors.[18] Thus, we find evidence that cognitive dissonance indeed backfires among the highly educated, both in a regression with a stratified sample of interest and in a regression using interaction terms among the full sample, suggesting a potential role for heterogeneous treatment of individuals to maximize e-billing uptake.

5. Discussion

In line with the literature, we find that environmental information does not affect individuals' propensity to opt into receiving paperless communications, even among purportedly green consumers. However, appealing to customers' desire for consistency of self-concept holds

[18] Calculation of the odds ratio for the effect of cognitive dissonance on the 'educated' sample: exp(0.048-0.342)=exp(-0.294)=0.745.

promise, though it backfires among our sample of individuals titled 'Doctor' or 'Professor'. Furthermore, our findings contradict the general conclusion in the literature that females are more likely to engage in environmental behaviors than males. The results indicate that informational campaigns are likely ineffective in promoting environmental behaviors, and that individuals with revealed altruistic preferences may be susceptible to messaging invoking feelings of cognitive dissonance. Imagery does not encourage environmental behavior in this context.

Given that the information provided is both easily available and free to access, the non-effect of environmental information speaks to many existing and emerging strands of literature on information and behavior. For example, the results fall in line with the notion of information avoidance, where individuals actively choose to evade information that might make them engage in altruistic behaviors that they otherwise do not wish to perform (Cain and Dana, 2012; Golman, Hagmann, and Loewenstein, 2015). An alternative explanation stemming from a phenomenon called moral licensing suggests that individuals who 'do good' along one dimension may allow themselves to 'do bad' (or simply not 'do good') along another (see Merritt, Effron, and Monin, 2010). Alternatively, perhaps the information is sufficient to change beliefs and intentions, though intentions have only been shown to be poorly correlated with behavior change (Webb and Sheeran, 2006). Another possible explanation could be that GE customers are already well aware of such information so that additional information has little effect on their beliefs—in line with a 'diminishing returns' argument (Stern, 2000)—or that the externalities are not sufficiently severe to induce change.

Moreover, our experiment demonstrates that particular individuals may be more or less susceptible to certain behavioral anomalies. In our case, individuals titled 'Doctor' or 'Professor' are far less likely to opt into e-billing if they receive the dissonance-inducing intervention as opposed to the control intervention. While education or status may be at play, we conjecture that this contrasting effect may be due to such individuals' altruistic fulfillment in their field of work. Therefore, issues of convenience—as highlighted in the control letter—may override concerns for maintaining an altruistic self-concept.

In sum, we recommend that green businesses abandon the use of information regarding environmental externalities as a tool to encourage environmentally beneficial decision making, and rather appeal to their customer base using more subtle tactics rooted in the psychology of cognitive dissonance, with careful attention to the audience of the messaging. Of course, there

are many additional tactics that could be equally—or possibly more—effective in encouraging particular types of customers to continue to make decisions in line with their past behavior. We note that this particular tactic may well generalize to other groups of socially responsible consumers, such as donors to particular causes or voters who have historically engaged in altruistic or civic behaviors. Further research should aim to gain a more nuanced understanding of the types of individual who may or may not be susceptible to messaging that appeals to desires for consistency in the self-concept.

REFERENCES

Abrahamse, Wokje, Linda Steg, Charles Vlek, and Talib Rothengatter. 2005. "A review of intervention studies aimed at household energy conservation." *Journal of Environmental Psychology* 25: 273-291.

Ajzen, Icek, and Martin Fishbein. 1980. *Understanding Attitudes and Predicting Social Behaviour.* Englewood Cliffs, NJ: Prentice-Hall.

Akerlof, George A., and William T. Dickens. 1982. "The economic consequences of cognitive dissonance." *American Economic Review* 72 (3): 307-319.

Akerlof, George A., and Rachel E. Kranton. 2005. "Identity and the economics of organizations." *The Journal of Economic Perspectives* 19 (1): 9-32.

Alfnes, Frode, Chengyan Yue, and Helen H. Jensen. 2010. "Cognitive dissonance as a means of reducing hypothetical bias." *European Review of Agricultural Economics* 37 (2): 147-163.

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95 (9): 1082-1095.

Andreoni, James. 1989. "Giving with impure altruism: Applications to charity and Ricardian equivalence." *The Journal of Political Economy* 97 (6): 1447-1458.

Andreoni, James. 1990. "Impure altruism and donations to public goods: A theory of warm-glow giving." *The Economic Journal* 100 (401): 464-477.

Aronson, Elliot. 1992. "The return of the repressed: Dissonance theory makes a comeback." *Psychological Inquiry* 3 (4): 303-311.

Axsom, Danny, and Joel Cooper. 1995. "Cognitive dissonance and psychotherapy: The role of effort justification in inducing weight loss." *Journal of Experimental Social Psychology* 21 (2): 149-160.

Bateson, Melissa, Daniel Nettle, and Gilbert Roberts. 2006. "Cues of being watched enhance cooperation in a real-world setting." *Biology Letters* 2: 412-414.

Becker, Gary. 1974. "A theory of social interactions." *National Bureau of Economic Research* Working Paper.

Burnham, Terence C., and Brian Hare. 2007. "Engineering human nature: Does involuntary neural activation increase public goods contributions?" *Human Nature* 18: 88-108.

Cain, Daylian M., and Jason Dana. 2012. "Paying people to look at the consequences of their actions." Working Paper.

Cheng, Tania, Danielle Kathryn Woon, and Jennifer K. Lynes. 2011. "The use of message framing in the promotion of environmentally sustainable behaviors." *Social Marketing Quarterly* 7 (2): 48-62.

Choi, James J., David Laibson, and Brigitte C. Madrian. 2005. "Are empowerment and education enough? Underdiversification in 401(k) plans." *Brookings Papers on Economic Activity* 2: 151-213.

Choi, James J., David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2002. "Defined contribution pensions: Plan rules, participant choices, and the path of least resistance." In *Tax Policy and the Economy, Volume 16*, pp. 67-114. MIT Press.

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

Dickerson, Chris Ann, Ruth Thibodeau, Elliot Aronson, and Dayna Miller. 1992. "Using cognitive dissonance to encourage water conservation." *Journal of Applied Social Psychology* 22 (11): 841-854.

Dijkstra, Arie, Hein De Vries, and Jolanda Roijackers. 1998. "Computerized tailored feedback to change cognitive determinants of smoking: A Dutch field experiment." *Health Education Research* 13 (2): 197-206.

Downs, Julie S., George Loewenstein, and Jessica Wisdom. 2009. "Strategies for promoting healthier food choices." *American Economic Review: Papers and Proceedings* 99 (2): 1-10.

Ekström, Mathias. 2012. "Do watching eyes affect charitable giving? Evidence from a field experiment." *Experimental Economics* 15: 530-546.

Elster, Jon. 1989. "Social norms and economic theory." *The Journal of Economic Perspectives* 3 (4): 99-117.

Eriksen, W., K. Serrum, and D. Bruusgaard. 1996. "Effects of information on smoking behaviour in families with preschool children." *Acta Paediatrica* 85 (2): 209-212.

Ernest-Jones, Max, Daniel Nettle, and Melissa Bateson. 2011. "Effects of eye images on everyday cooperative behavior: A field experiment." *Evolution and Human Behavior* 32: 172-178.

Ferraro, Paul J. and Michael K. Price. 2013. "Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment." *Review of Economics and Statistics* 95 (1): 64-73.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Stanford: Stanford University Press.

Gilad, Benjamin, Stanley Kaish, and Peter D. Loeb. 1987. "Cognitive dissonance and utility maximization: A general framework." *Journal of Economic Behavior and Organization* 8 (1): 61-73.

Goetzmann, William N., and Nadav Peles. 1997. "Cognitive dissonance and mutual fund investors." *Journal of Financial Research* 20 (2): 145-158.

Goldsmith, Arthur H., Stanley Sedo, William Darity Jr., and Darrick Hamilton. 2004. "The labor supply consequences of perceptions of employer discrimination during search and on-the-job: Integrating neoclassical theory and cognitive dissonance." *Journal of Economic Psychology* 25: 15-39.

Golman, Russell, David Hagmann, and George Loewenstein. 2015. "Information avoidance." Available at SSRN 2633226.

Goodright, Victoria, and Emily Wilkes. 2015. "Chapter 3: Domestic energy consumption in the UK between 1970 and 2014." in *Energy Consumption in the UK (2015).*

Haley, Kevin J. and Daniel M. T. Fessler. 2005. "Nobody's watching? Subtle cues affect generosity in an anonymous economic game." *Evolution and Human Behavior* 26: 245-256.

Harnack, Lisa J., Simone A. French, J. Michael Oakes, Mary T. Story, Robert W. Jeffery, and Sarah A. Rydell. 2008. "Effects of calorie labeling and value size pricing on fast food meal choices: Results from an experimental trial." *International Journal of Behavioral Nutrition and Physical Activity* 5 (1): 63-76.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect theory: An analysis of decision under risk." *Econometrica* 47 (2): 263-292.

Kollmuss, Anja, and Julian Agyeman. **2002.** "Mind the gap: Why do people act environmentally and what are the barriers to pro-environmental behavior?" *Environmental Education Research* 8 (3): 239-260.

Konow, James. 2000. "Fair shares: Accountability and cognitive dissonance in allocation decisions." *American Economic Review* 90 (4): 1072-1091.

List, John A., Sally Sadoff, and Mathis Wagner. 2011. "So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design." *Experimental Economics* 14 (4): 439-457.

Madrian, Brigitte C., and Dennis F. Shea. 2001. "Preaching to the converted and converting those taught: Financial education in the workplace." University of Chicago Working Paper.

Mannberg, Andréa. 2012. "Risk and rationalization: The role of affect and cognitive dissonance for sexual risk taking." *European Economic Review* 56 (6): 1325-1337.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The dishonesty of honest people: A theory of self-concept maintenance." *Journal of Marketing Research* 45 (6): 633-644.

Merritt, Anna C., Daniel A. Effron, and Benoît Monin. 2010. "Moral self-licensing: When being good frees us to be bad." *Social and Personality Psychology Compass* 4/5: 344-357.

Mifune, Nobuhiro, Hirofumi Hashimoto, and Toshio Yamagishi. 2010. "Altruism toward in-group members as a reputation mechanism." *Evolution and Human Behavior* 31: 109-117.

Mullainathan, Sendhil, and Ebonya Washington. 2006. "Sticking with your vote: Cognitive dissonance and voting." *National Bureau of Economic Research* Working Paper No. w11910.

Nail, Paul R., Julia E. Misak, and Randi M. Davis. 2004. "Self-affirmation versus self-consistency: A comparison of two competing self-theories of dissonance phenomena." *Personality and Individual Differences* 36 (8): 1893-1905.

Oxoby, Robert J. 2003. "Attitudes and allocations: Status, cognitive dissonance, and the manipulation of attitudes." *Journal of Economic Behavior and Organization* 52 (3): 365-385.

Rabin, Matthew. 1994. "Cognitive dissonance and social change." *Journal of Economic Behavior and Organization* 23 (2): 177-194.

Rennekamp, Kristina, Kathy K. Rupar, and Nicholas Seybert. 2015. "Impaired judgment: The effects of asset impairment reversibility and cognitive dissonance on future investment." *The Accounting Review* 90 (2): 739-759.

Rigdon, Mary, Keiko Ishii, Motoki Watabe, and Shinobu Kitayama. 2009. "Minimal social cues in the dictator game." *Journal of Economic Psychology* 30: 358-367.

Schwartz, Shalom H. 1973. "Normative explanations of helping behavior: A critique, proposal, and empirical test." *Journal of Experimental Social Psychology* 9: 349-364.

Sexton, Steven. 2015. "Automatic bill payment and salience effects: Evidence from electricity consumption." *Review of Economics and Statistics* 97 (2): 229-241.

Staats, H. J., A. P. Wit, and C. Y. H. Midden. 1996. "Communicating the greenhouse effect to the public: Evaluation of a mass media campaign from a social dilemma perspective." *Journal of Environmental Management* 45: 189-203.

Stern, Paul C. 2000. "Toward a coherent theory of environmentally significant behavior." *Journal of Social Issues* 56 (3): 407-424.

Stern, Paul C., Thomas Dietz, Troy D. Abel, Gregory A. Guagnano, and Linda Kalof. 1999. "A value-belief-norm theory of support for social movements: The case of environmentalism." *Human Ecology Review* 6 (2): 81-97.

Taylor, Shelley E. and Suzanne C. Thompson. 1982. "Stalking the elusive 'vividness' effect." *Psychological Review* 89 (2): 155-181.

Toledo, Chantal. 2016. "Do environmental messages work on the poor? Experimental evidence from Brazilian favelas." *Journal of the Association of Environmental and Resource Economists* 3 (1): 37-83.

Vohs, Kathleen D. 2015. "Money priming can change people's thoughts, feelings, emotions, and behaviors: An update on 10 years of experiments." *Journal of Experimental Psychology* 144 (4): e86-e93.

Webb, Thomas L., and Paschal Sheeran. 2006. "Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence." *Psychological Bulletin* 132 (2): 249-268.

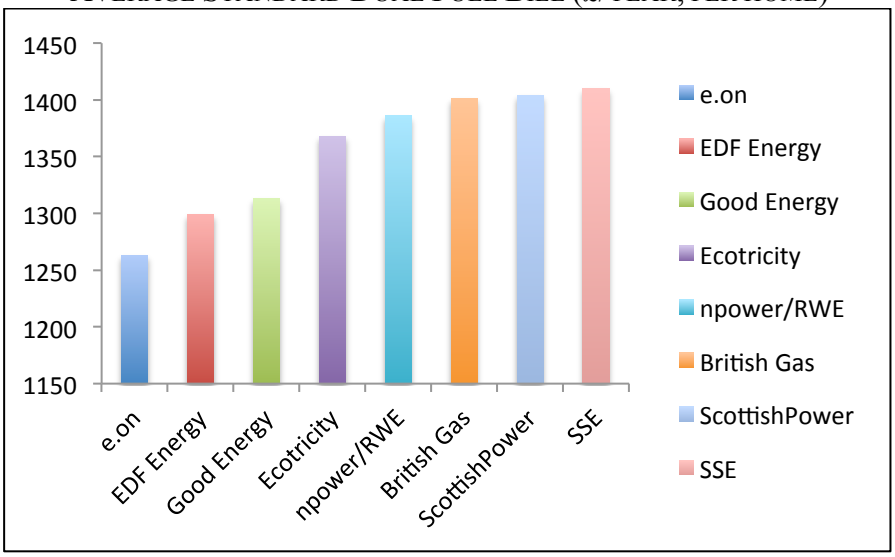
Wichardt, Philipp C. 2012. "Norms, cognitive dissonance, and cooperative behaviour in laboratory experiments." *International Journal of Social Economics* 39 (5): 342-356.

Zelezny, Lynette C. 1999. "Educational interventions that improve environmental behaviors: A meta-analysis." *The Journal of Environmental Education* 31 (1): 5-14.

FIGURES AND TABLES

FIGURE 1

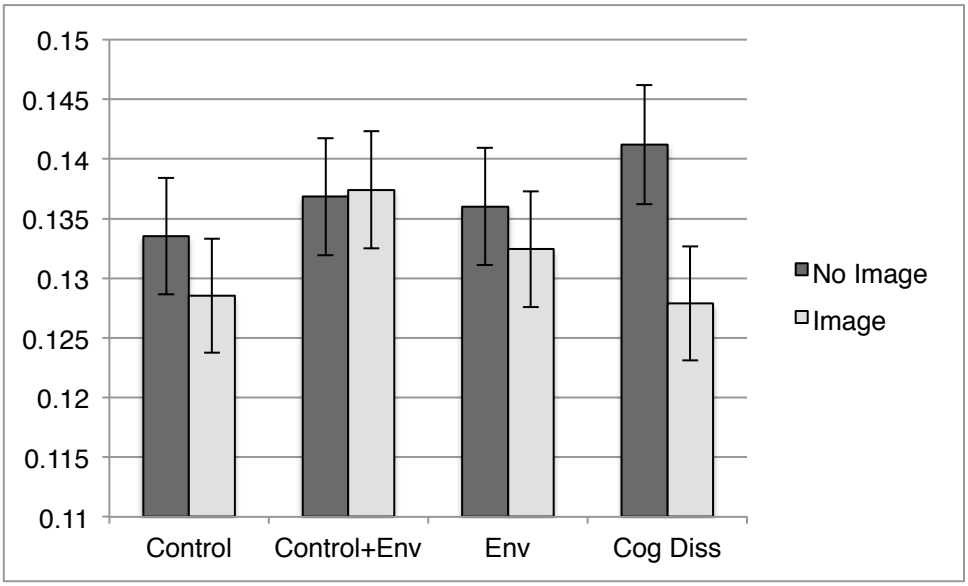
AVERAGE STANDARD DUAL FUEL BILL (£/YEAR, PER HOME)



Notes: The data above was taken from Energy Helpline on 18 November 2013 and is based on 3,300 kWh of electricity and 16,500 kWh of natural gas paid using direct debit on the standard variable rate. The source of this chart is "Green Energy Suppliers in the UK Compared to the Big 6", accessed 30 March 2016 < http://shrinkthatfootprint.com/green-electrical-supply-uk-big-6>.

FIGURE 2

E-BILLING UPTAKE ACCORDING TO GROUP ASSIGNMENT



Notes: The above bar graph shows the proportion of each study group that signed up to e-billing, with standard error bars.

TABLE 1
EXPERIMENTAL DESIGN

Content	Text	Study Groups			
		Control *(Groups 1-2)*	*Environmental Framing* *(Groups 3-4)*	*Control + Environmental Framing* *(Groups 5-6)*	*Cognitive Dissonance* *(Groups 7-8)*
Availability and Online Access	**It's finally here!** Now you can switch to e-billing and have your energy bills emailed directly to your inbox rather than receiving them by post.	✓	✓	✓	✓
Customer Benefits	**The benefits of switching from paper billing to e-billing:** • Access bills 24/7 online ; • Spend less time sorting through mail; • Reduce paper waste;	✓		✓	✓
Environmental Benefits	**If all customers make the switch, we would save 46 trees worth of paper each year!** **Why reduce paper waste?** • The average UK family throws away 6 trees worth of paper in their household bin each year. • Paper production ranks 3rd and 4th for most energy intensive and greenhouse gas intensive manufacturing industries (respectively). • 12.5 million tonnes of paper and cardboard are used annually in the UK, making us the 11th worst paper offender in the world.		✓	✓	
Environmental Steward	**As a Good Energy customer, you are an environmental steward.** By switching to e-billing, you take another important step to eliminate the environmental impact of your energy use.				✓

Notes: While the 'Control and Environmental Framing' intervention simply adds environmental information to the Control email, the email doubles in length with the addition. Therefore, we also include the 'Environmental Framing' intervention that is a similar length and format to the Control email so that we can 'control' for the added complexity of including a large amount of additional information to the Control email. All even-numbered groups receive the treatment with the image.

TABLE 2A
BALANCE CHECK: GROUPS WITH VS. WITHOUT IMAGES

	Group 1	Group 2	Test of equality: G1=G2	Group 3	Group 4	Test of Equality: G3=G4	Group 5	Group 6	Test of Equality: G5=G6	Group 7	Group 8	Test of Equality: G7=G8
Fuel Type:												
Dual Fuel	0.409	0.411	p=0.824	0.408	0.407	p=0.991	0.410	0.409	p=0.930	0.409	0.409	p=0.975
	(0.492)	(0.492)		(0.491)	(0.491)		(0.492)	(0.492)		(0.492)	(0.492)	
Gas	0.062	0.062	p=0.949	0.062	0.062	p=0.967	0.063	0.061	p=0.628	0.062	0.062	p=0.943
	(0.241)	(0.241)		(0.241)	(0.241)		(0.243)	(0.239)		(0.242)	(0.241)	
Electricity	0.529	0.527	p=0.803	0.530	0.531	p=0.975	0.527	0.530	p=0.748	0.529	0.530	p=0.948
	(0.499)	(0.499)		(0.499)	(0.499)		(0.499)	(0.499)		(0.499)	(0.499)	
Gas Consumption	13.949	13.807	p=0.602	13.633	13.781	p=0.575	13.863	13.605	p=0.324	13.886	13.672	p=0.426
	(9.352)	(9.025)		(8.757)	(9.092)		(9.038)	(8.590)		(9.284)	(8.864)	
Electricity Consumption	3.720	3.622	p=0.190	3.625	3.672	p=0.531	3.753	3.626	p=0.107	3.685	3.640	p=0.548
	(3.845)	(3.231)		(3.419)	(3.671)		(4.162)	(3.283)		(3.419)	(3.615)	
Days as Customer	314.8	313.9	p=0.887	317.4	317.2	p=0.977	312.1	313.3	p=0.861	316.7	318.8	p=0.770
	(333.9)	(321.0)		(338.3)	(344.0)		(327.5)	(333.8)		(342.4)	(346.7)	
Gender	0.469	0.468	p=0.952	0.470	0.470	p=0.991	0.470	0.471	p=0.907	0.469	0.472	p=0.834
	(0.499)	(0.499)		(0.499)	(0.499)		(0.499)	(0.499)		(0.499)	(0.499)	
Postgraduate Education	0.045	0.046	p=0.860	0.048	0.051	p=0.446	0.050	0.051	p=0.830	0.046	0.045	p=0.677
	(0.208)	(0.210)		(0.213)	(0.220)		(0.217)	(0.219)		(0.210)	(0.207)	
Day of Week			p=0.777			p=0.846			p=0.962			p=0.983
Sample size	4817	4825		4834	4850		4830	4838		4825	4836	

Notes: The table checks for balance across observables for groups with identical intervention content, where one group receives the environmental image and the other does not. The p-values in the table derive from chi-square tests (for comparisons of dummy and categorical variables) and t-tests (for comparisons of continuous variables). Group 1 is the Control group, 2 is Control with image, 3 is the Control and Environmental Framing, 4 is Control and Environmental Framing with image, 5 is Environmental Framing, 6 is Environmental Framing with image, 7 is Cognitive Dissonance, and 8 is Cognitive Dissonance with image. The table pertains to individuals in the entire sample, except for the following: gender balance tests are conducted only for individuals for whom gender is identified, and balance tests on annual gas and electricity consumption are conducted only for individuals who consume gas and energy, respectively. Annual estimated energy and gas consumption are measured at the unit of 1000 kWh. The *fuel type* dummy variables specify the type of fuel the customer receives from Good Energy, where "dual fuel" indicates that they receive both gas and electricity. Gas and electricity consumption are estimated annual usage values measured at the unit of 1000 kWh. Female is equal to one if the customer is female, and postgraduate education is equal to 1 if the customer holds a title of 'Doctor' or 'Professor'. Day of week is a categorical variable indicating the day of week on which the customer received the treatment email; since means do not provide valuable information for this variable, we simply report the p-value for the chi-square test. Standard deviations are reported below means in parentheses.

TABLE 2B

BALANCE CHECK: CONTROL VS. TREATMENTS

Groups	Test of Equality: G1=G3	Test of Equality: G1=G4	Test of Equality: G1=G5	Test of Equality: G1=G6	Test of Equality: G1=G7	Test of Equality: G1=G8
Fuel Type:						
Dual Fuel	p=0.886	p=0.877	p=0.923	p=0.993	p=0.998	p=0.971
Gas	p=0.934	p=0.968	p=0.762	p=0.856	p=0.947	p=0.972
Electricity	p=0.919	p=0.894	p=0.809	p=0.937	p=0.972	p=0.985
Gas Consumption	p=0.240	p=0.540	p=0.751	p=0.197	p=0.819	p=0.305
Electricity Consumption	p=0.214	p=0.540	p=0.700	p=0.208	p=0.645	p=0.310
Days as Customer	p=0.709	p=0.733	p=0.681	p=0.815	p=0.787	p=0.573
Gender	p=0.897	p=0.906	p=0.937	p=0.845	p=0.949	p=0.792
Postgraduate Education	p=0.622	p=0.210	p=0.330	p=0.234	p=0.820	p=0.850
Day of Week	p=0.925	p=0.992	p=0.998	p=0.971	p=0.912	p=0.759

Notes: The table checks for balance on observables between the control group and all treatment groups (see Table 1A for means and sample sizes). The p-values in the table derive from chi-square tests (for comparisons of dummy and categorical variables) and t-tests (for comparisons of continuous variables); see Table. Group 1 is the Control group, 2 is Control with image, 3 is the Control and Environmental Framing, 4 is Control and Environmental Framing with image, 5 is Environmental Framing, 6 is Environmental Framing with image, 7 is Cognitive Dissonance, and 8 is Cognitive Dissonance with image. The table pertains to individuals in the entire sample, except for the following: gender balance tests are conducted only for individuals for whom gender is identified, and balance tests on annual gas and electricity consumption are conducted only for individuals who consume gas and energy, respectively. Annual estimated energy and gas consumption are measured at the unit of 1000 kWh. The *fuel type* dummy variables specify the type of fuel the customer receives from Good Energy, where "dual fuel" indicates that they receive both gas and electricity. Gas and electricity consumption are estimated annual usage values measured at the unit of 1000 kWh. Female is equal to one if the customer is female, and postgraduate education is equal to 1 if the customer holds a title of 'Doctor' or 'Professor'. Day of week is a categorical variable indicating the day of week on which the customer received the treatment email; since means do not provide valuable information for this variable, we simply report the p-value for the chi-square test.

TABLE 3

LOGIT REGRESSION – MAIN SAMPLE

	OR	Marginal	OR	Marginal
G2: Control, Image	0.971	-0.003	0.968	-0.004
	(0.060)	(0.007)	(0.060)	(0.007)
G3: Control Env	1.017	0.002	1.018	0.002
	(0.062)	(0.007)	(0.063)	(0.007)
G4: Control Env, Image	0.997	-0.000	0.996	-0.000
	(0.061)	(0.007)	(0.062)	(0.007)
G5: Env	1.042	0.005	1.042	0.005
	(0.064)	(0.007)	(0.064)	(0.007)
G6: Env, Image	1.046	0.005	1.047	0.005
	(0.064)	(0.007)	(0.064)	(0.007)
G7: Cog Diss	1.105*	0.012*	1.107*	0.012*
	(0.067)	(0.007)	(0.067)	(0.007)
G8: Cog Diss, Image	0.964	-0.004	0.965	-0.004
	(0.060)	(0.007)	(0.060)	(0.007)
Gas Consumption			0.996*	-0.001*
			(0.002)	(0.000)
Energy Consumption			0.986***	-0.002***
			(0.005)	(0.001)
Tariff: Gas Only			0.597***	-0.050***
			(0.043)	(0.006)
Tariff: Electric Only			0.569***	-0.065***
			(0.026)	(0.005)
Female			0.735***	-0.035***
			(0.023)	(0.003)
Constant	0.152***		0.257***	
	(0.007)		(0.015)	
Observations	36,810	36,810	36,810	36,810
Controls	No	No	Yes	Yes

Notes: The above logit regression pertains to individuals in the main sample. Annual estimated energy and gas consumption are measured at the unit of 1000 kWh.

TABLE 4

LOGIT REGRESSION – POSTGRADUATE EDUCATION SAMPLE

	OR	Marginal	OR	Marginal
G2: Control, Image	0.946	-0.007	0.929	-0.009
	(0.245)	(0.032)	(0.242)	(0.032)
G3: Control Env	1.110	0.014	1.100	0.012
	(0.275)	(0.034)	(0.274)	(0.033)
G4: Control Env, Image	0.992	-0.001	0.986	-0.002
	(0.249)	(0.032)	(0.249)	(0.032)
G5: Env	0.943	-0.007	0.936	-0.008
	(0.243)	(0.032)	(0.242)	(0.032)
G6: Env, Image	0.867	-0.018	0.869	-0.017
	(0.222)	(0.031)	(0.224)	(0.030)
G7: Cog Diss	0.582*	-0.060**	0.570*	-0.061**
	(0.166)	(0.027)	(0.164)	(0.026)
G8: Cog Diss, Image	0.917	-0.011	0.876	-0.016
	(0.241)	(0.032)	(0.232)	(0.031)
Gas Consumption			1.001	0.000
			(0.009)	(0.001)
Energy Consumption			0.969	-0.004
			(0.024)	(0.003)
Tariff: Gas Only			0.446***	-0.080***
			(0.137)	(0.023)
Tariff: Electric Only			0.655**	-0.054**
			(0.132)	(0.026)
Constant	0.197***		0.285***	
	(0.036)		(0.070)	
Observations	1,844	1,844	1,844	1,844
Controls	No	No	Yes	Yes

Notes: The above logit regression pertains to individuals in the main sample. Annual estimated energy and gas consumption are measured at the unit of 1000 kWh.

APPENDIX

TABLE A1
PROPORTION SIGNED UP TO E-BILLING:
T-TESTS COMPARING EXPERIMENTAL CONDITIONS

	Control (C)	Environmental Framing (EF)	Test of Equality: C vs. EF	Control + Environmental Framing (CEF)	Test of Equality: C vs. CEF	Test of Equality: EF vs. CEF	Cognitive Dissonance (CD)	Test of Equality: C vs. CD
No Image	0.134	0.136	p=0.539	0.138	p=0.693	p=0.826	0.142	p=0.226
	(0.340)	(0.343)		(0.345)			(0.349)	
	N=4817	N=4830		N=4834			N=4824	
Image	0.130	0.134	p=0.624	0.138	p=0.261	p=0.526	0.129	p=0.893
	(0.337)	(0.340)		(0.345)			(0.335)	
	N=4825	N=4838		N=4850			N=4836	
Pooled	0.132	0.135	p=0.532	0.138	p=0.219	p=0.546	0.136	p=0.439
	(0.338)	(0.342)		(0.345)			(0.343)	
	N=9642	N=9668		N=9684			N=9660	

Notes: The table shows the results of tests of equality of means (t-tests) for rate of sign-up across experimental conditions for all subjects in the study sample, where groups with and without images (e.g., G1 and G2) are pooled in the final row. Standard deviations are presented below means in parentheses.

TABLE A2
EFFECTS OF CONTROL VARIABLES
(CONTROL GROUP ONLY)

	OR	Marginal
Electricity Consumption	0.993	-0.001
	(0.007)	(0.001)
Gas Consumption	0.985	-0.002
	(0.014)	(0.002)
Tariff: Gas Only	0.581***	-0.051***
	(0.118)	(0.016)
Tariff: Electricity Only	0.530***	-0.072***
	(0.069)	(0.015)
Female	0.735***	-0.034***
	(0.065)	(0.010)
Constant	0.272***	
	(0.032)	
Observations	4,598	

Notes: The above logit regression pertains to the individuals in the control group (without image) of the main sample. Annual estimated energy and gas consumption are measured at the unit of 1000 kWh.

TABLE A3

GENDER AND TREATMENT

	OR	Marginal
G2: Control, Image	0.942	-0.007
	(0.077)	(0.009)
G3: Control Env	1.091	0.010
	(0.087)	(0.010)
G4: Control Env, Image	0.976	-0.003
	(0.080)	(0.009)
G5: Env	0.969	-0.004
	(0.079)	(0.009)
G6: Env, Image	1.090	0.010
	(0.087)	(0.010)
G7: Cog Diss	1.118	0.013
	(0.089)	(0.010)
G8: Cog Diss, Image	0.972	-0.003
	(0.079)	(0.009)
G2*Female	1.069	0.008
	(0.135)	(0.015)
G3*Female	1.182	0.020
	(0.146)	(0.016)
G4*Female	0.906	-0.011
	(0.113)	(0.013)
G5*Female	0.842	-0.018
	(0.106)	(0.013)
G6*Female	1.049	0.006
	(0.131)	(0.015)
G7*Female	0.977	-0.003
	(0.120)	(0.014)
G8*Female	0.982	-0.002
	(0.124)	(0.014)
Gas Consumption	0.996*	-0.001*
	(0.002)	(0.000)
Energy Consumption	0.986***	-0.002***
	(0.005)	(0.001)
Tariff: Gas Only	0.596***	-0.050***
	(0.043)	(0.006)
Tariff: Electric Only	0.569***	-0.065***
	(0.026)	(0.005)
Female	0.737***	-0.034***
	(0.065)	(0.010)
Constant	0.256***	
	(0.018)	
Observations	36,810	

Notes: The above logit regression pertains to the individuals in the main sample. Annual estimated energy and gas consumption are measured at the unit of 1000 kWh.

TABLE A4

POSTGRADUATE EDUCATION AND TREATMENT

	OR	Marginal
G2: Control, Image	0.969	-0.004
	(0.060)	(0.007)
G3: Control Env	1.018	0.002
	(0.063)	(0.007)
G4: Control Env, Image	0.996	-0.000
	(0.062)	(0.007)
G5: Env	1.042	0.005
	(0.064)	(0.007)
G6: Env, Image	1.047	0.005
	(0.064)	(0.007)
G7: Cog Diss	1.107*	0.012
	(0.067)	(0.007)
G8: Cog Diss, Image	0.965	-0.004
	(0.060)	(0.007)
G2*Educ	0.956	-0.005
	(0.256)	(0.030)
G3*Educ	0.898	-0.012
	(0.238)	(0.028)
G4*Educ	0.826	-0.020
	(0.219)	(0.026)
G5*Educ	1.074	0.008
	(0.276)	(0.031)
G6*Educ	0.987	-0.002
	(0.256)	(0.030)
G7*Educ	0.513**	-0.060***
	(0.150)	(0.020)
G8*Educ	0.912	-0.010
	(0.247)	(0.029)
Gas Consumption	0.996	-0.000
	(0.002)	(0.000)
Energy Consumption	0.987***	-0.002***
	(0.005)	(0.001)
Tariff: Gas Only	0.605***	-0.049***
	(0.042)	(0.006)
Tariff: Electric Only	0.581***	-0.063***
	(0.026)	(0.005)
Educ	1.319	0.035
	(0.248)	(0.026)
Constant	0.220***	
	(0.012)	
Observations	38,654	

Notes: The above logit regression includes all individuals in the study sample. Annual estimated energy and gas consumption are measured at the unit of 1000 kWh.

FIGURE A1
CONTROL INTERVENTION

From: **Good Energy** noreply@goodenergy.co.uk
Subject: Go paperless with Good Energy!
Date: September 4, 2014 at 5:58 AM
To:

GE

Good Energy

Switch for Good

Dear

It's finally here! Now you can switch to e-billing and have your energy bills emailed directly to your inbox rather than receiving them by post.

Even better, you can access your bills online any time, so they won't fill any valuable space in your drawers or bins.

Here at Good Energy, we prioritise customer satisfaction. The opportunity to switch to e-billing is just one more step we have taken to keep you smiling.

The benefits of switching from paper billing to e-billing:

- Reduce paper waste
- Spend less time sorting through mail
- Access bills 24/7 online

Go on – it's easy! **Switch to e-billing here**.

Let's work together to better the world of energy.

Best wishes,

Dave Ford
Chief Operating Officer

Good Energy, Monkton Reach, Monkton Hill, Chippenham, SN15 1EE
Registered Office: Good Energy Limited, Monkton Reach, Monkton Hill, Chippenham, SN15 1EE
Company Registration No. 3899612, Place of Registration: England and Wales. VAT No. 811 3295 57

Notes: This e-mail (and any attachments) may be confidential and may contain personal views which are not the views of Good Energy Limited unless specifically stated. If you have received it in error, please delete it from your system, do not use, copy or disclose the information in any way nor act in reliance on it and notify the sender immediately. Please note that Good Energy Limited monitors e-mails sent or received. Further communication will signify your consent to this.

goodenergy.co.uk Contact us Facebook Twitter Blog

FIGURE A2
ENVIRONMENTAL FRAMING INTERVENTION

From: **Good Energy** noreply@goodenergy.co.uk
Subject: Go paperless with Good Energy!
Date: September 4, 2014 at 7:55 AM
To:

GE

Good Energy

Switch for Good

Dear

It's finally here! Now you can switch to e-billing and have your energy bills emailed directly to your inbox rather than receiving them by post.

If all of our customers make the switch, we would save 46 trees worth of paper each year!*

Even better, you can access your bills online any time, so they won't fill any valuable space in your drawers or bins.

Here at Good Energy, we prioritise customer satisfaction as well as the environment. The opportunity to switch to e-billing is just one more step we have taken to keep you smiling and help you shrink your environmental footprint.

Why reduce paper waste?

- The average UK family throws away 6 trees worth of paper in their household bin each year.
- Paper production ranks 3rd and 4th for most energy intensive and greenhouse gas intensive manufacturing industries (respectively).
- 12.5 million tonnes of paper and cardboard are used annually in the UK, making us the 11th worst paper offender in the world.

Go on – it's easy! **Switch to e-billing here**.

Let's work together to better the world of energy.

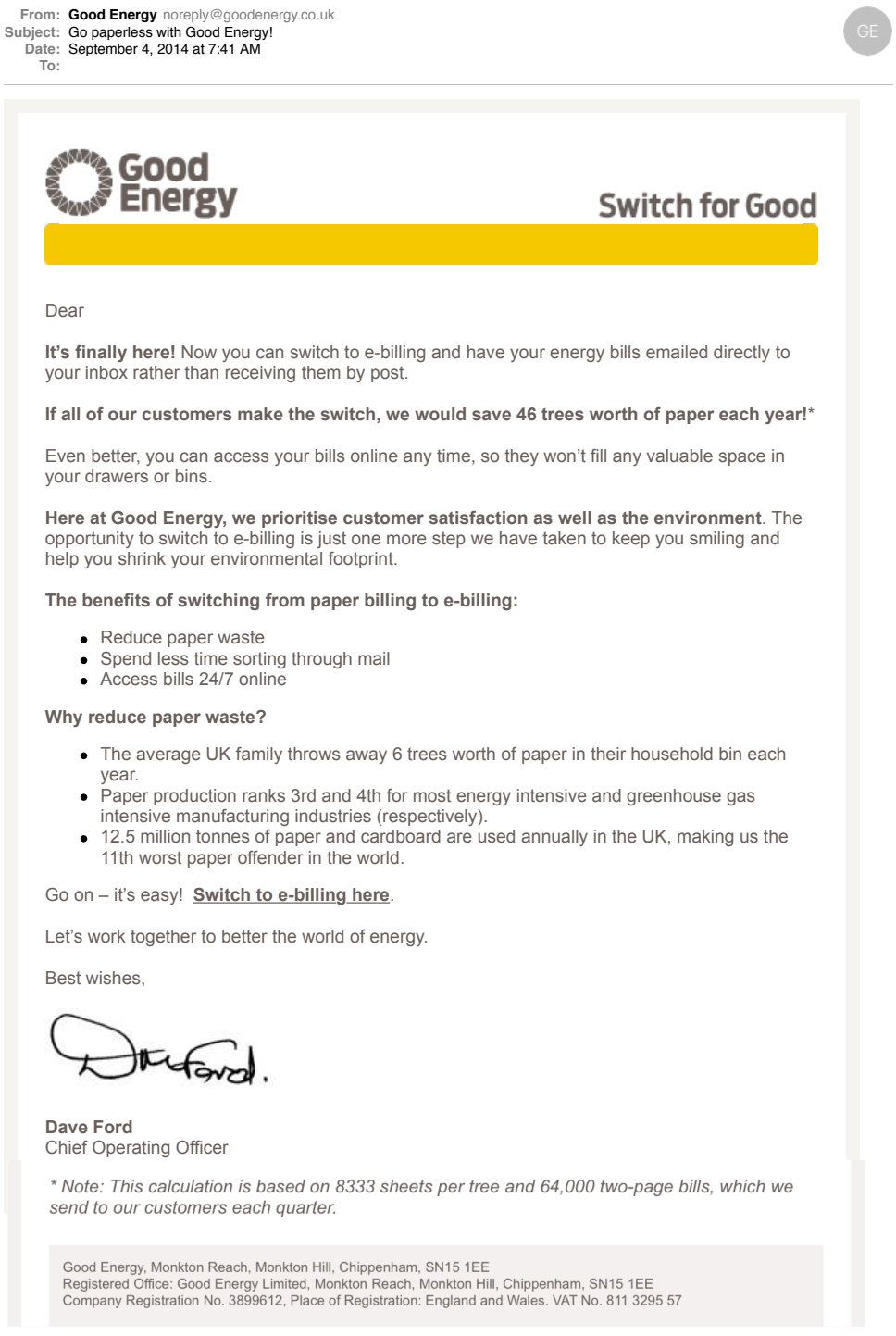
Best wishes,

Dave Ford
Chief Operating Officer

** Note: This calculation is based on 8333 sheets per tree and 64,000 two-page bills, which we send to our customers each quarter.*

Good Energy, Monkton Reach, Monkton Hill, Chippenham, SN15 1EE
Registered Office: Good Energy Limited, Monkton Reach, Monkton Hill, Chippenham, SN15 1EE
Company Registration No. 3899612, Place of Registration: England and Wales. VAT No. 811 3295 57

FIGURE A3
CONTROL AND ENVIRONMENTAL FRAMING INTERVENTION



From: **Good Energy** noreply@goodenergy.co.uk
Subject: Go paperless with Good Energy!
Date: September 4, 2014 at 7:41 AM
To:

GE

Good Energy

Switch for Good

Dear

It's finally here! Now you can switch to e-billing and have your energy bills emailed directly to your inbox rather than receiving them by post.

If all of our customers make the switch, we would save 46 trees worth of paper each year!*

Even better, you can access your bills online any time, so they won't fill any valuable space in your drawers or bins.

Here at Good Energy, we prioritise customer satisfaction as well as the environment. The opportunity to switch to e-billing is just one more step we have taken to keep you smiling and help you shrink your environmental footprint.

The benefits of switching from paper billing to e-billing:

- Reduce paper waste
- Spend less time sorting through mail
- Access bills 24/7 online

Why reduce paper waste?

- The average UK family throws away 6 trees worth of paper in their household bin each year.
- Paper production ranks 3rd and 4th for most energy intensive and greenhouse gas intensive manufacturing industries (respectively).
- 12.5 million tonnes of paper and cardboard are used annually in the UK, making us the 11th worst paper offender in the world.

Go on – it's easy! **Switch to e-billing here**.

Let's work together to better the world of energy.

Best wishes,

Dave Ford
Chief Operating Officer

** Note: This calculation is based on 8333 sheets per tree and 64,000 two-page bills, which we send to our customers each quarter.*

Good Energy, Monkton Reach, Monkton Hill, Chippenham, SN15 1EE
Registered Office: Good Energy Limited, Monkton Reach, Monkton Hill, Chippenham, SN15 1EE
Company Registration No. 3899612, Place of Registration: England and Wales. VAT No. 811 3295 57

FIGURE A4
COGNITIVE DISSONANCE INTERVENTION

From: **Good Energy** noreply@goodenergy.co.uk
Subject: Go paperless with Good Energy!
Date: September 4, 2014 at 8:02 AM
To:

GE

Good Energy

Switch for Good

Dear

It's finally here! Now you can switch to e-billing and have your energy bills emailed directly to your inbox rather than receiving them by post.

Even better, you can access your bills online any time, so they won't fill any valuable space in your drawers or bins.

As a Good Energy customer, you are an environmental steward. By switching to e-billing, you take another important step to eliminate the environmental impact of your energy use.

The benefits of switching from paper billing to e-billing:

- Access bills 24/7 online
- Spend less time sorting through mail
- Reduce paper waste

Go on – it's easy! **Switch to e-billing here**.

Let's work together to better the world of energy.

Best wishes,

Dave Ford
Chief Operating Officer

Good Energy, Monkton Reach, Monkton Hill, Chippenham, SN15 1EE
Registered Office: Good Energy Limited, Monkton Reach, Monkton Hill, Chippenham, SN15 1EE
Company Registration No. 3899612, Place of Registration: England and Wales. VAT No. 811 3295 57

Notes: This e-mail (and any attachments) may be confidential and may contain personal views which are not the views of Good Energy Limited unless specifically stated. If you have received it in error, please delete it from your system, do not use, copy or disclose the information in any way nor act in reliance on it and notify the sender immediately. Please note that Good Energy Limited monitors e-mails sent or received. Further communication will signify your consent to this.

goodenergy.co.uk Contact us Facebook Twitter Blog

FIGURE A5
EMAIL IMAGE

CHAPTER III

A NEW APPROACH TO AN AGE-OLD PROBLEM: SOLVING EXTERNALITIES BY INCENTING WORKERS DIRECTLY

By Greer Gosnell, John List, and Robert Metcalfe

Abstract: Understanding motivations in the workplace remains of utmost import as economies around the world rely on increases in labor productivity to foster sustainable economic growth. This study makes use of a unique opportunity to "look under the hood" of an organization that critically relies on worker effort and performance. By partnering with Virgin Atlantic Airways on a field experiment that includes over 40,000 unique flights covering an eight-month period, we explore how information and incentives affect captains' performance. Making use of more than 110,000 captain-level observations, we find that our set of treatments—which include performance information, personal targets, and prosocial incentives—induces captains to improve efficiency in all three key flight areas: pre-flight, in-flight, and post-flight. We estimate that our treatments saved between 266-704 tons of fuel for the airline over the eight-month experimental period. These savings led to between 838-2,200 tons of CO_2 abated at a marginal abatement cost of *negative* $250 per ton of CO_2 (i.e. a $250 savings per ton abated). Methodologically, our approach highlights the potential usefulness of moving beyond an experimental design that focuses on short-run substitution effects, and it also suggests a new way to combat firm-level externalities: target workers rather than the firm as a whole.

Acknowledgments: We thank participants at the 2015 EEE and PPE NBER Summer Institute sessions for excellent remarks that considerably improved the research, and seminar participants at Columbia University, Dartmouth College, University of British Columbia, University of Chicago, University of Tennessee, and University of Wisconsin-Madison. Omar Al-Ubaydli, Steve Cicala, Diane Coyle, Paul Dolan, Robert Dur, Robert Hahn, Glenn Harrison, Justine Hastings, David Jimenez-Gomez, Matthew Kahn, Kory Kroft, Edward Lazear, Steve Levitt, Bentley MacLoed, Jonathan Meer, Michael Norton, Sally Sadoff, Laura Schechter, Jessie Shapiro, Kathryn Shaw, Kerry Smith, Alex Teytelboym, Gernot Wagner, and Catherine Wolfram provided remarks that helped to sharpen our thoughts. Thanks to The Templeton Foundation and the Science of Philanthropy Initiative at the University of Chicago for providing the generous funds to make this experiment possible. Further thanks to the UK Civil Aviation Authority and the pilots' unions who took the time to review and approve of the study objectives and material. A

special thanks to those individuals at Virgin Atlantic Airways—especially to Paul Morris, Claire Lambert, Emma Harvey, and Captain David Kistruck—and Rolls Royce (especially Mark Goodhind and Simon Mayes) for their essential roles in the implementation of this experiment. These parties are in no way responsible for the analyses and interpretations presented in this paper. We thank Florian Rundhammer and Andrew Simon for their excellent research assistance.

1. Introduction

Many scientists believe that global climate change represents the most pervasive externality of our time (Stern, 2007). Perhaps one of the lowest-hanging fruits in combating climate change is to design firm-level incentive schemes for workers to engage in green behaviors. Given the Environmental Protection Agency estimate that 21 percent of carbon emissions in the United States are from firms (U.S. Environmental Protection Agency, 2015), there is undoubtedly much to gain. Yet, very few studies have explored incentive aspects within the workplace that pertain to sustainability, whether it is shifting work hours to less energy-intensive times of the day or incenting employees to use fewer resources per unit of output.[19] Indeed, when resource use is linked to production costs (as is almost always the case), mitigating the externality has the potential to foster increased profits, providing distinct possibilities of a win-win scenario.

Consider the transportation sector, and in particular air transportation of humans and cargo. The airline industry is a significant contributor to human welfare, with over three billion passengers per year and 35% of the value of world trade transported by air (Federal Aviation Administration, 2015). However, the global aviation industry is directly responsible for significant health costs among vulnerable population groups (Schlenker and Walker, 2016).[20] Moreover, excessive fuel use in the industry affects profits—fuel represents an average 33% of airlines' operating costs (Air Transport Action Group, 2014)—and poses a severe risk to the global environment. Emissions from the air transport sector currently account for 3.5-5% of global radiative forcing and 2-3% of global carbon dioxide emissions (Penner et al., 1999; Lee et al., 2009; Burkhardt and Kärcher, 2011), deeming the industry a significant force in climate change discussions.[21]

Technology adoption and market-based instruments continue to appear on the industry's agenda as primary means to reach its dual goals of carbon neutral growth by 2020 and halving

[19] Atkin et al. (2015) demonstrate low adoption of waste-reducing technology among soccer ball producers in Pakistan, demonstrating that incentives to use the technology increase uptake. Freeman and Kleiner (2005) study the use of incentive pay on production costs, finding that piece rate wages may increase individual productivity, though not enough to offset the costs associated with monitoring and requisite managerial policies.

[20] Schlenker and Walker (2016) focus on the effects of network delays in the east coast of the United States on congestion at large airports in California to assess health effects from daily variation in air pollution. These effects are presumed to be generalizable across large airports globally and are a consequence of the aviation industry as a whole.

[21] Past research has shown that the airline industry has also not fully internalized social costs associated with crashes (Borenstein and Zimmerman, 1988). Here we highlight yet another means by which the social cost of the industry is not incorporated into its decision calculus. Nonetheless, demand for air travel is forecasted to increase over the next two decades and, as a result, airline emissions will likely trend upwards (Borenstein, 2011).

greenhouse gas emissions from 2005 levels by 2050 (International Civil Aviation Organization, 2013). Yet, despite large potential to reduce fuel burn from eliminating operational inefficiencies (Green, 2009; Singh and Sharma, 2015), almost no research has been undertaken to understand the potential for cost and emissions savings from changes in the behavior of transport personnel. In fact, we are unaware of research, more generally, on the optimal incentive structure for employees to engage in conservation activities in the workplace.

Our study takes a strong initial step toward such an understanding by partnering with Virgin Atlantic Airways (VAA) on a field experiment. We observe over 40,000 unique flights over a 27-month period for the entire population of captains eligible to fly both before and during the experiment.[22] In the aviation industry, airline captains maintain a considerable amount of autonomy when it comes to fuel and flight decisions. We capitalize on recent technological developments that capture detailed flight-level data to measure captains' fuel efficiency across three distinct phases—pre-flight, in-flight, and post-flight.[23] The pre-flight measure (denoted *Fuel Load*) assesses the accuracy with which captains implement final adjustments to aircraft fuel load given all relevant factors (e.g., weather and aircraft weight).[24] The in-flight measure (denoted *Efficient Flight*) assesses how fuel-efficiently the captain operates the aircraft between takeoff and landing. The post-flight measure (denoted *Efficient Taxi*) provides information on how fuel-efficiently the captain operates the aircraft once on the ground. The experiment explores the extent to which several experimental treatments—implemented from February 2014 through September 2014—influence captains' behaviors.[25]

[22] The "captain"—as opposed to the "first officer"—is the pilot on the aircraft who makes command decisions and is ultimately responsible for the flight's safety. As a rule, captains are the most senior pilots in an airline (see Smith (2013) for insight into captains' roles and responsibilities). In the cockpit of a typical flight from New York to London, there would be one captain and one first officer on board who both engage (more or less equally) in aircraft operations, though the captain is ultimately responsible for all aspects of flight operation. A vast majority of airline captains survive rigorous job market competition to secure their jobs, investing thousands of hours of training (privately or elsewhere) before obtaining the opportunity to be considered for a flying career with a major airline. A handful of VAA captains who were on leave for personal reasons or were fulfilling duties outside of their usual obligations were excluded from the sample.

[23] The Fuel Efficiency team within Virgin Atlantic Airways was responsible for identification of the fuel-efficient behaviors targeted in this study, which represent the outcomes of just a few of the many decisions that a captain engages with during a given flight.

[24] Captains do not have much in the way of decision support tools for calculating the correct Fuel Load apart from pen and a receipt-like sheet of paper—printed in the cockpit prior to departure—indicating the final weight of the aircraft. They then use pen and paper to make two calculations using a rule of thumb that first prescribes the amount of additional fuel to load for the flight and subsequently dictates the additional fuel necessary to carry added fuel.

[25] Captains were assured on several occasions that their participation in the experiment held no implications or consequences for their salaries or career prospects. For instance, the initial letter sent to all (treatment and control) captains in January 2014 included the following statements (emphasis included): *"**This is not, in any way, shape or form, an attempt to set up a 'fuel league table', or any attempt at moving in the direction of a fuel league table.** It is an independent research project to see whether information provided in different ways affects individual decisions. All data gathered during this study will remain **anonymous** and **confidential...** Again, we would like to stress that Captains' anonymity will be maintained throughout the study; whilst somebody in Flight Ops Admin has to correlate which Captain gets which letter, Flight Operations Management will have no visibility of which Captain is in which Group, and who is doing what in response to which information. Information will be sent to all Captains in the active study groups. What you choose to do with that information is entirely up to you."*

The treatments are inspired by a simple principal-agent model wherein we attempt to influence the behaviors of VAA captains. Our theoretical model yields predictions on how the act of measurement itself might yield behavioral change, in the spirit of the Hawthorne effects described in Levitt and List (2011). In addition, the model shows how performance information, personal targets, and prosocial incentives for reaching those targets can motivate behavioral change. As such, our experimental design revolves around understanding how the act of measurement itself as well as each of these three factors—information about recent fuel efficiency, exogenous targets, and prosocial incentives (a donation to the captain's chosen charity conditional on achieving the target provided)—affect captains' behaviors from pre-flight to post-flight. The present study is the first to evaluate the separate elements of performance-related pay (PRP) schemes in a high-stakes setting with experienced professional workers. Many other studies test base pay versus PRP; however, PRP has three distinct behavioral elements that may drive a change in behavior: informational feedback, a conditional target, and the incentive itself. Here, these elements are broken down and distributed across treatment conditions.

We are unaware of any previous research that tests the impacts of targets or prosocial incentives on worker productivity in a high-stakes professional setting. Moreover, it should be noted that the present context is not a typical principal-agent setting in which the principal does not observe effort. Here, the principal has accurate measures of effort. However, since the highly unionized labor force holds significant bargaining power, the principal faces restrictions against contracting on effort—or on output, for that matter, which is the typical contracting variable in a basic principal-agent model. Therefore, the firm is in a "second-best" world of needing to use behavioral incentives instead of the financial incentives in PRP.

Making use of more than 110,000 observations of behavior across 335 captains, we find several interesting insights that have the potential to alter conventional approaches to motivating employee effort in the workplace while reducing both operating costs and environmental damage. Perhaps most surprisingly, by simply informing captains that we—i.e. the academic researchers and VAA Fuel Efficiency personnel overseeing the study—are measuring their behaviors on

three dimensions, we are able to considerably reduce fuel inefficiency.[26] For example, captains in the control group significantly increased the implementation of Efficient Flight and Efficient Taxi by nearly 50 percent from the pre-experimental period. These behavioral changes generated more than 6,800 tons of fuel saved for the airline over the eight-month experimental period (i.e. $5.37 million in fuel savings), which translates to more than 21,000 tons of CO_2 abated.

Despite these large Hawthorne effects, we find a significant role for the three experimental treatments. The information treatment increases effort for Efficient Taxi, but does not increase effort for Fuel Load or Efficient Flight. We find, however, that personal targets increase effort for Efficient Flight and Efficient Taxi. Finally, prosocial incentives increase effort across all three dimensions. Furthermore, we find significant differences between information and the two treatments that provide targets, while we do not detect differential effects between the target and prosocial treatment groups. That is, adding conditional prosocial incentives in the form of a donation to the captain's chosen charity does not provide further lift beyond the effects of a personal target.[27] Yet, there is an interesting effect of prosocial incentives: they induce a reduction in flight time by an average of 1 minute and 30 seconds per flight relative to the control group, equivalent to more than 80 hours of reduced flight time over the course of the study.[28]

The difference-in-difference treatment effect estimates indicate that the various interventions increased implementation of fuel-efficient activities by 1-10 percentage points above the pre-experimental period (i.e. additional to the Hawthorne effect).[29] Based on these effects, we estimate that the three treatments saved between 266-704 tons ($209,000-$553,000) of fuel for the airline over the eight-month experimental period. This fuel savings corresponds to 838-2,220 tons of CO_2 abated. Since the cost of the treatments is merely the cost of postage

[26] This pure monitoring effect aligns with agency theory (e.g., Alchian and Demsetz, 1972; Stiglitz, 1975), as well as with experimental results such as those in Boly (2011). These results are also related to the work of Hubbard (2000, 2003), who found that monitoring truckers' performance using GPS technology leads to improved performance for those workers where driver effort is important and where verifying drivers' actions to insurers is valuable. He estimates that such technology has increased capacity utilization by around 3% in the trucking industry. VAA policies precluded the designation of an uninformed control group, so estimates of Hawthorne effects are based on before-and-after comparisons, as in Bandiera, Barankay, and Rasul (2007, 2009). Nonetheless, our results suggest that the data before the experiment was stationary and there was an upward trend once the experiment started. Importantly, since all information provided to captains in treatment groups is individual-specific, we are able to rule out contamination (i.e. spillover effects of information) as a possible contributor to the change in behavior exhibited by the control group.

[27] To our knowledge, we are the first to experimentally estimate the impact of an incentive given to a charity if the worker reaches a certain performance target in his or her job. Our notion of prosocial incentives is therefore different to the social incentives presented in Bandiera et al. (2009, 2010), who demonstrate the manner in which social connections in the workplace influence workers' and managers' motivations.

[28] These results are based on all flights and are presented in Table A4. The total flight time reduction is calculated by multiplying the average effect of captains in the prosocial treatment group relative to control by the number of flights undertaken in the prosocial treatment group during the study period.

[29] Since there were no upward trends before the experiment began—that is, a Dickey-Fuller test indicates that pre-treatment behaviors were stationary—we can be confident that the experiment improved fuel efficiency from business as usual.

materials (here, $855 per treatment group), the marginal abatement cost (MAC) of the treatments is minuscule, falling between $1.02 and $0.39 per ton of carbon saved. However, since the airline benefited from significant cost savings via reduced fuel usage as a result of the interventions, the MAC in this context is approximately -$250 per ton in actuality (using 2014 prices). Such an astonishingly low MAC outperforms every other reported carbon abatement technology of which we are aware (see Enkvist, Nauclér, and Rosander, 2007; McKinsey, 2009).[30]

Our experimental design highlights the usefulness of moving beyond short-run effects in favor of understanding long-term embedded behavior change. First, in terms of persistence of the treatment effects throughout the experiment, we find that the largest effects for Fuel Load and Efficient Flight arise in the middle months of the experiment, while the treatment effect for Efficient Taxi is consistently high throughout. Interestingly, the largest effects occur on the behavior that is the easiest to change (Efficient Taxi). Once the experiment finishes, however, we find that captains' effort reverts to post-experiment baseline levels (i.e. equivalent attainment to the control group once the experiment terminates) for Fuel Load and Efficient Flight, while the treatment effects attenuate for Efficient Taxi. With regards to the persistence of the Hawthorne effect, the post-experiment baseline remains considerably improved from the pre-experiment baseline, indicating that monitoring induces captains to make low-effort efficiency improvements that are quickly and easily habituated. An alternative interpretation to the Hawthorne effect could be that the captains now learn that the firm values fuel efficiency (a value that captains likely share). This interpretation relates to the work by Bloom and Van Reenen (2007) and Bloom et al. (2014) on the impact of "soft" management styles and structures on worker productivity.

Our findings are relevant for academics, businesses, and policymakers alike. For academics, the theory and experimental results hold implications for environmental, behavioral, labor, and public economics. For example, there exist movements within both applied economics ("X-efficiency"; see Leibenstein, 1966) and environmental economics (the "Porter Hypothesis"; see Porter and van der Linde, 1995) arguing that substantial "free gains" exist within firms. The premise is a behavioral one: rather than modeling firms as fully aware and understanding of all extant means to maximize resource efficiency—thereby exhausting all cost-efficient measures at each moment in time—this approach considers the firm as a composition of networks of

[30] The most cost-effective abatement strategy according to McKinsey (2009) is switching residential lighting from incandescent bulbs to LED bulbs at a MAC of approximately -165 Euros, or about -$177.

boundedly rational individuals burdened by problematic principal-agent incentive conflicts (Leibenstein, 1966; Perelman, 2011). To support this view, a survey of evidence argues that a typical firm operates at 65% to 97% efficiency (Button and Weyman-Jones, 1992), though much of this evidence is based on observational data and does not assess impacts based on a true counterfactual. Our work complements this environmental and behavioral research by moving in a hitherto unconsidered direction: rather than focus on capital improvements or research and development, we explore efficiency effects of incenting labor directly during their normal course of work. For labor economists considering principal-agent settings, our study suggests that allowing the agent flexibility to achieve goals might be a key trigger in enhancing effort profiles.

For businesses and policymakers, we present a novel and promising approach to combating firm-level externalities: design appropriate incentives for workers. More narrowly, the study provides practical and cost-effective fuel solutions for the air transport industry. Our empirical approach lends itself naturally to related tests across other sectors of the economy. By making use of our theoretical framework to guide experimental treatments in the field, businesses and policymakers can learn not only *what* works, but also *why* it works. This understanding will provide decision makers with a more effective toolkit to advance efficient policies and procedures.

The remainder of the paper is structured as follows. Section 2 provides contextual background, a sketch of the theory of captain behavior, and the experimental design. Section 3 presents the experimental results. Section 4 provides a discussion related to policy implications and related avenues for future research.

2. Background, theory, and experimental design

In 2012, we began discussions with VAA to partner on a field experiment with the aim of understanding behavioral components of fuel usage without adversely affecting safety practices or job satisfaction.[31] We developed a theoretical framework and a field experiment (detailed further below) that allowed us to remain within institutional constraints while maintaining the integrity of lending theoretical insights to the experimental data. We agreed to provide monthly tailored feedback to 335 airline captains—the entire eligible captain population of VAA—from

[31] The study is a component of Change is in the Air, VAA's wider sustainability initiative (see http://www.virgin-atlantic.com/content/dam/VAA/Documents/sustainabilitypdf/SustainabilityPolicy201407.05.14.pdf).

February 2014 through September 2014. Importantly, *all* eligible VAA captains were included in the experiment—in either control or treatment. These captains have absolute authority to make all fuel-related decisions. They range in experience and fly long-haul flights on various aircraft types (Airbus 330-300, Airbus 340-300, Airbus 340-600, Boeing 744-400, Boeing 787-9). We include a map of destinations in Figure 1.[32]

While many of the captains' choices are important in terms of fuel efficiency outcomes, VAA identified three primary measurable levers to change behavior for the purpose of this study. The first lever is a pre-flight consideration, which VAA refers to as the Zero Fuel Weight (ZFW) adjustment. Approximately 90 minutes before each flight, captains utilize flight-specific flight plan information (e.g., expected fuel usage, weather, and aircraft weight) in conjunction with their own professional judgment to determine initial fuel uptake, which usually corresponds to approximately 90% of the anticipated fuel necessary for the flight. This amount is fueled into the aircraft simultaneous to the loading of passengers and cargo. Near to completion of passenger boarding and cargo/baggage loading, the pilots—now on the flight deck—receive updated information regarding the final weight of the aircraft and may adjust the fuel on the aircraft accordingly. The information they receive from Flight Operations includes a ZFW measure, which indicates the weight of the aircraft with the revenue load (i.e. passengers and cargo), as well as the Takeoff Weight (TOW), which includes both revenue load and fuel.

Captains then perform a ZFW calculation in which they first calculate the amount by which they should increase or decrease fuel load based on the final ZFW—a formula that is standard across the airline industry. If they have decided to increase the fuel load, they subsequently compute a second iteration to account for the additional fuel necessary to carry the fuel that they have decided to add to the aircraft. If the amount of fuel already on the aircraft is sufficient according to these calculations, the captain may choose not to add any additional fuel.

For mnemonic purposes, rather than use ZFW we denote this binary outcome variable as Fuel Load. Fuel Load indicates whether the double iteration calculation has been performed and the fuel level adjusted accordingly.[33] We deem the captains' behavior successful if their final

[32] All operations to Aberdeen and Edinburgh are VAA Little Red operations (i.e. branded VAA flights operated by a third party) and were excluded from the analysis. In April 2015, VAA removed its service to Cape Town; this route change took place subsequent to the period covered in our dataset.

[33] Using data from an anonymous major U.S. airline, Ryerson et al. (2015) estimate that 4.5% of fuel burned on an average flight is attributable to carrying unused fuel, and that more than 1% of fuel burned on an average flight is due to addition of contingency fuel "above a reasonable buffer".

fuel load is within 200 kg of the "correct" amount of fuel as dictated by the calculation. This allowance prevents penalizing captains for rounding and slight over- or under-fueling on the part of the fueler while providing measurable targets for captains in two of our treatment groups. According to our partner airline, accurate Fuel Load adjustment should ideally be performed on every flight regardless of circumstances, which would correspond to 100% attainment for the performance metric provided.

The second lever is an in-flight consideration: Efficient Flight. The Efficient Flight metric captures whether captains (and their co-pilots) use less fuel during flight than is allotted in the updated flight plan.[34] We use this metric to understand whether captains have made fuel-efficient choices between takeoff and landing. It incorporates several in-flight behaviors that augment fuel efficiency, such as requesting and executing optimal altitudes and shortcuts from air traffic control, maintaining ideal speeds, optimally adjusting to en route weather updates, and ensuring efficient aerodynamic arrangements with respect to flap settings as well as takeoff and landing gear. The Efficient Flight metric affords captains the flexibility to achieve the target while using professional judgment to ensure that safety remains the first priority. Under some uncommon circumstances, operational requirements dictate that captains sacrifice fuel efficiency (and VAA accepts the captains' decisions as final), so we would not expect even a "model" captain to perform this metric on 100% of flights, though the metric should be attainable on a vast majority of flights. In our analysis, Efficient Flight equals 1 if the captain does not exceed the projected fuel use for that flight (adjusted for actual TOW), and 0 otherwise.[35]

The final lever—reduced-engine taxi-in (Efficient Taxi or Efficient Taxiing, hereafter)—occurs post-flight. Once the aircraft has landed and the engines have cooled, captains may choose to shut down one (or two, in a four-engine aircraft) of their engines while they taxi to the gate, thereby decreasing fuel burn per minute spent taxiing. Captains meet the criteria for this metric if they shut down (at least) one engine during taxi in.[36] As with Efficient Flight, there are circumstances under which the airline would not expect or prescribe the implementation of Efficient Taxi. Obstacles include geographical constraints (e.g., the placement or layout of the runway) and the complexity of the taxi route (e.g., number of stops, turns, or cul-de-sacs). Still,

[34] The flight plan is updated subsequent to decisions made on Fuel Load so that decisions regarding the first metric do not affect one's ability to meet this in-flight metric.
[35] Note that it was essential to create binary metrics for Fuel Load and Efficient Flight so we could assign targets to captains in the targets and prosocial incentives group.
[36] Fuel savings from Efficient Taxi depend on scheduling and delays as savings are accrued on a per-minute basis. Savings also depend on aircraft type and only begin to accrue after engines have cooled, which takes 2-5 minutes from touch down. Savings per minute for aircraft operated within our study are as follows: 12.5 kg (B744, A330), 8.75 kg (A346), and 6.25 kg (A343).

the metric should also be attainable on a vast majority of flights, and obstacles to implementation are uncorrelated with treatment.

Fuel Load, Efficient Flight, and Efficient Taxi are the three primary outcome variables in the experiment. It is important to recognize that fuel is a major cost to airlines—accounting for roughly 33 percent[37] of total operating costs—and has been rising over the last fifteen years (Borenstein, 2011; International Air Transport Association, 2014). Thus, airlines are interested in cost-effective means to reduce fuel burn. Given their renowned expertise and experience in the industry, however, airline captains are granted significant autonomy in their decision making across several fuel-relevant behaviors, including those described above. Moreover, captains' strong unionization makes it contractually difficult to use performance-related pay to induce efficiency.[38] As such, we focus on alternative motivations to reduce fuel usage.

i. Theoretical sketch of captains' behavior

We model an airline captain's choices using a static game of a principal-agent model that determines a captain's chosen effort in a given period (for parsimony, we briefly sketch the model here and provide details in Appendix II). The tasks consist of the aforementioned pre-flight to post-flight fuel usage metrics. Captains observe their own effort and a signal of optimal fuel usage; the signal is noisy unless the captain receives information. Captains' perspectives on fuel usage and their fuel-relevant decisions are rooted in their own experiences and preferences and are conditional on contextual (i.e. flight- and day-specific) factors.

Captains choose how much effort to exert to maximize a utility function that includes utility from wealth, job performance, and charitable giving, as well as disutility from effort exertion and social pressure. The model has the standard prediction from the first-order conditions that the captain will expand effort until its marginal cost equals the marginal utility gained from the associated decrease in fuel usage. This prediction occurs on several dimensions, such as utility from job performance, utility from giving to a charity, as well as disutility from social pressure (a la DellaVigna, List, and Malmendier, 2012).

[37] For the airline represented in this study, fuel accounts for 35% of operating costs.
[38] For a discussion of how unionization can affect the long-run outcomes of firms, see Lee and Mas (2012).

Although our base model follows DellaVigna et al. (2012), we extend the model to incorporate a reference-dependent component to capture the effects of exogenous targets. In line with existing theories of reference dependence, we posit that a change in one's personal expectations from the status quo to an improved outcome can boost performance and, consequently, utility. We therefore introduce feedback to employees providing non-binding targets—i.e. focal points for attainment of the three fuel-relevant behaviors—that encapsulate reference-dependent preferences.[39] We expect utility from job satisfaction to increase for those who meet their targets. As in the Köszegi and Rabin (2006) model of reference-dependent preferences, we assume individuals are loss averse so that performing below the target level will cause more disutility than exceeding the target level will benefit the individual.

The notion that prosocial incentives can motivate behavior change is rooted in theories of pure and impure altruism (Becker, 1974; Andreoni, 1989, 1990). Pure altruism requires that individuals derive utility from the benefits they directly receive from the provision of a public good. Impure altruism posits that individuals gain utility from the act of giving itself, so that an individual whose altruism is completely impure will provide the same dollar value toward the public good regardless of the provision of others. Both pure and impure altruism provide positive utility to (altruistic) economic agents, and we assume that individuals are characterized by some combination of the two (we do not attempt to distinguish between them in our experiment). This characterization provides a prediction that altruistic motivations combined with charitable incentives will augment fuel efficiency.

In equilibrium, captains choose the corresponding effort level that satisfies the first-order conditions. These choices lead to several propositions.[40] First, if social pressure is important, then captains in the control group will improve their fuel efficiency due to the enhanced scrutiny of their fuel usage. Second, providing information to captains will cause them to increase (weakly decrease) their effort if estimated fuel usage is lower (higher) than their actual fuel usage. The intuition is that the relationship between captains' estimated fuel usage and their actual fuel usage importantly determines their utility from job performance. For example, informing captains that they are fuel-inefficient will induce captains to exert greater effort if they derive disutility from consuming more fuel than their estimated usage. Alternatively, if their fuel

[39] There is a rich psychology literature on goal setting. Heath, Larrick, and Wu (1999) present evidence that goals act as reference points inducing loss aversion and diminishing sensitivity in a manner consistent with Prospect Theory (see also Locke and Latham, 2006). Psychology studies do not exist that address the complex and high-stake field environment in which our experiment takes place.

[40] These propositions would remain unchanged if we set up the model in the vein of a multi-tasking model.

usage is deemed lower than the estimated usage, they might exert less effort since effort is costly.

Third, targets set above pre-study fuel use will cause captains to weakly increase their effort. Captains will increase their effort if the marginal gain from the associated decrease in fuel usage due to the target is greater than the marginal cost of effort. Alternatively, captains will not increase their effort if the marginal cost of effort is larger than the marginal gain from the associated decrease in fuel usage in the job performance parameter. Fourth, conditional donations to charity will increase effort if captains' altruism is strictly positive and will not affect their effort otherwise. Fifth, of the three dimensions to lower fuel usage—pre-flight, in-flight, and post-flight—captains will choose to increase their effort the most in tasks for which the targets are least costly to meet (i.e. Efficient Taxi).

In light of these predictions, we design a field experiment to measure how behaviors related to fuel usage are affected by: i) information about recent fuel efficiency, ii) information about target fuel efficiency, and iii) a donation to a chosen charity conditional on achieving the target efficiency. To our knowledge, we are the first to perform a large-scale field experiment on firm employees in a high-stakes professional labor setting (where the average salary of a captain is roughly \$175,000-\$225,000[41]).[42] In doing so, we overcome prominent labor market frictions in the airline industry by implementing interventions that do not change contracts of the captains.[43] We outline the field experimental design below.

[41]This salary range is based on information updated in June 2015: http://www.pilotjobsnetwork.com/jobs/Virgin_Atlantic.

[42] There is a growing literature surrounding field labor economics, but most experiments have focused on simple tasks (List and Rasul, 2011; Bandiera et al., 2011; Levitt and Neckermann, 2014). Using a before-and-after design within the same company, Bandiera et al. (2007, 2009, 2010) demonstrate the effects of managerial compensation and social connections in the workplace on worker productivity and selection in the fruit picking industry. Shearer (2004) finds that piece-rate wages improve worker productivity relative to fixed-rate wages in tree planting; Lazear (1999) finds similar incentive effects of piece-rate wages in an observational study of automobile glass installers. Field experiments on the impact of retail store-level tournaments on sales show mixed results (Delfgaauw et al., 2013, 2014, 2015), while a quasi-experiment showed that simply informing warehouse employees of relative wage standing permanently improved productivity (Blanes i Vidal and Nossol, 2011). One exception to such task simplicity is Gibbs, Neckermann, and Siemroth (2014), who analyze the effects of a rewards program on innovation at a large Asian technology firm in a field experimental setting. They find that providing rewards for idea acceptance substantially increases the quality of ideas submitted. In an envelope-stuffing experiment, Al-Ubaydli et al. (2015) find that quality is actually higher under piece-rate wages (contrary to predictions from economic theory), speculating a role for beliefs about employers' ability to monitor. In an artefactual field experiment with bicycle messengers, Burks et al. (2009) find that performance-related pay reduces cooperation in a prisoner's dilemma game relative to a flat wage. There has been some research by Rockoff et al. (2012) that demonstrates that simple information on teacher performance to employers can improve productivity in schools, increase turnover for teachers with low performance estimates, and produce small test score improvements.

[43] While a standard principal-agent model would prescribe the use of contracted performance-related pay to align captains' fuel use incentives with those of the airline, the airline workforce is a different labor market to most due to the high skill requirements (and often government safety certifications) necessary to enter this particular labor force. See Borenstein and Rose (2007) for a further discussion of the labor market frictions of the aviation industry.

ii. Experimental design

In accordance with our theoretical model, our field experiment focuses on three behavioral motivations for optimizing fuel use: personalized information, performance targets, and prosocial incentives. The three treatments centered upon three behaviors central to fuel use: Fuel Load, Efficient Flight, and Efficient Taxi. Respectively, these three behaviors allow us to capture captains' behavior before takeoff, during the flight, and after landing. Airline captains did not receive detailed information relating their decision making to their fuel efficiency prior to this experiment (consistent with both airline and industry standards). Recent advances in aircraft data collection allow us to obtain precise data to inform captains of the link between their effort and their efficiency.

We partnered with VAA's Sustainability and Fuel Efficiency teams to provide accurate monthly feedback to three treatment groups over the course of eight months across 335 captains; a control group did not receive any feedback but was aware that their fuel usage was being monitored.[44] Printed feedback reports with information from the previous month's flights were sent to the home addresses of treated captains, so that captains received their first feedback report in mid-March 2014 and their final feedback report in mid-October 2014. The three experimental treatments can be summarized as follows:

Treatment Group 1: Information. Each feedback report details the captain's performance of the three fuel-relevant behaviors for the prior month. Specifically, the feedback presents the percentage of flights flown during the preceding month for which the captain successfully implemented each of the three behaviors. For instance, if a captain flew four flights in the prior month, successfully performing Fuel Load and Efficient Taxi on two of the flights and Efficient Flight on three of the flights, his feedback report would indicate a 50% attainment level for the former behaviors and a 75% attainment level for the latter.

[44] In keeping with VAA's culture of transparency, carefully crafted study information sheets were posted to captains' home addresses on January 20, 2014. These information sheets guaranteed captains of the anonymity of their data and assured them that the study was not a step in the direction of competitive league tables. Additionally, captains in treatment groups received a notification of their assigned treatment group with a sample feedback form, including the appropriate targets for captains in Treatment Groups 2 and 3, which were posted on January 27, 2014, five days prior to the first day of monitoring. Since participants were aware that they were part of an experiment, our field experiment should be considered a framed field experiment in the parlance of Harrison and List (2004). Yet, unlike any other framed field experiment of which we are aware, we are estimating a parameter devoid of selection bias since all captains are experimental subjects. In this way, our behavioral parameter of interest shares much with that estimated in a natural field experiment (see Al-Ubaydli and List, 2015).

Treatment Group 2: Targets. Captains in this treatment group received the same information outlined above but were additionally encouraged to achieve personalized targets of 25% above their pre-experimental baseline attainment levels for each metric (capped at 90%). The targets were communicated to these captains prior to the start of the experiment. An additional box is included in the feedback report to provide a summary of performance (i.e. total number of targets met). If at least two of the three targets were met, captains were recognized with an injunctive statement ("*Well Done!*") and encouraged to continue to fly efficiently the following month. If fewer than two targets were met, captains were encouraged to fly more efficiently to reach their targets. Captains were not rewarded or recognized in any public or material fashion for their achievements.

Treatment Group 3: Prosocial Incentives. In addition to the information and targets provided to captains in treatment group 2, those in the prosocial treatment group were informed that achieving their targets would result in donations to charity. Specifically, for each target achieved in a given month, £10 was donated to a charity of the captains' choice on their behalves.[45] Therefore, captains in this treatment group each had the opportunity to donate £30 ($49) per month for a total of £240 ($389) to their chosen charity over the course of the eight-month trial. Captains were reminded each month of the remaining potential donations that could result from realizing their targets in the future. To our knowledge, ours is the first randomized field study to use performance-based charitable incentives to increase employee effort.[46] Table 1 outlines the treatments (see Appendix III for examples of each of the three feedback reports).

This "build-on" design allows us to assess whether there are additional benefits of prosocial incentives beyond sole provision of information and personal targets, the latter of which have an extremely low marginal cost to the principal.[47] Within our experiment, we did not change any organizational structures or contracts with the airline captains, although we recognize

[45] When captains in the prosocial treatment group were informed of their assignment to treatment, they were offered the opportunity to choose one of five diverse charities to support with their charitable incentives: Free the Children, MyClimate, Help for Heroes, Make A Wish UK, and Cancer Research UK. Eighteen captains selected a charity by emailing the designated project email address, and 67 captains who did not actively select a charity were defaulted to donate to Free the Children. Captains could choose to remain anonymous; otherwise, exact donations were attributed to each individual (identified by their first initial and last name).

[46] See Imas (2014) and Charness, Cobo-Reyes, and Sanchez (2014) for lab experiments on the effect of charitable incentives on effort, and Anik et al. (2013) for a field study of unconditional charitable bonuses. Relatedly, field experimental research into unconditional gifts is a burgeoning area of research—see Gneezy and List (2006); Bellemare and Shearer (2009); Hennig-Schmidt et al. (2010); Englmaier and Leider (2012); Kube, Maréchal, and Puppe (2012); and Cohn, Fehr, and Goette (2015).

[47] The closest research to this "free lunch" approach is depicted in the field experiments of Grant and Gino (2010); Kosfeld and Neckermann (2011); Bradler et al. (2013); Chandler and Kapelner (2013); Gubler, Larkin, and Pearce (2013); Ashraf, Bandiera, and Jack (2014); Ashraf, Bandiera, and Lee (2014); Kosfeld, Neckermann, and Yang (2014).

that these could be important to productivity and efficiency.[48] Importantly, our design uses incentive schemes that permit flexibility for workers to achieve their goals. In this way, rather than mandate or incent a particular course of action, we follow a more adaptable approach that permits gains to be had in accord with the captains' personal and professional discretion.

iii. Additional experimental details

Randomization. To randomize subjects across the four groups, the pre-experimental data (September-November, 2013) were first blocked on five dummy variables that captured whether subjects were above or below average for: i) number of engines on aircraft flown, ii) number of flights executed per month, and iii) attainment for the three selected fuel-relevant behaviors. The former two variables were those that proved significant in determining the selected outcome behaviors in preliminary regressions, while the three target behaviors are our main dependent variables. Once blocked, subjects in each block were randomly allocated to one of the four study groups through a matched quadruplet design. To ensure that individual-specific observable characteristics are balanced across groups, we performed balance tests for gender, seniority, age, trainer status, and whether the captain participated in the selective pre-study focus group. In addition to checking for balance across the variables on which the data were blocked, we checked for balance on flight plan fuel (i.e. as a proxy for average flight distance). In short, an exploration of all available aspects of captain and flight data reveals that the randomization was successful in that the observables are balanced across the four experimental conditions (see Tables 2 and 3).

Communication with captains. Two weeks prior to the beginning of the study, all captains were informed that VAA would be undertaking a study on fuel efficiency as part of its Change is in the Air sustainability initiative. The initial letter outlined the behaviors to be measured and the possible study groups to which the captains may be assigned. Captains in treatment groups were to receive letters the following week to inform them of what to expect in the coming months. In the final week of January 2014, letters were sent to all treated captains informing them of the intervention to which they had been assigned. The letter contained a sample feedback report including targets, if applicable.

[48] See Nagin et al. (2002); Hamilton, Nickerson, and Owan (2003); Karlan and Valdivia (2011); Bandiera et al. (2013); Bloom et al. (2013); Karlan, Knight, and Udry (2015); Bloom et al. (2015). Our context is the single firm experimental setting in the insider econometrics approach (Shaw, 2009).

From February 1, 2014 to October 1, 2014, we gathered all flight-level data on a monthly basis for each captain and mailed a feedback report to the home address of each treated captain.[49] Captains were encouraged to engage with the material and send any questions to an email address created specifically for study inquiries. Once the experiment was complete, we sent treated captains a debrief letter informing them of their overall monthly results with respect to their targets (if in the targets or prosocial treatment groups) and their total charitable donations (if in the prosocial treatment group). All (treatment and control) captains were informed that a follow-up survey would be sent to their company email addresses in early 2015.[50]

Sample. Our data consist of the entire eligible universe of VAA captains (N=335), of which 329 are male and 6 are female. Of the debrief survey respondents, 97 classified their training as military and 102 as civilian (the remaining declined to state). Eleven captains are "trusted pilots" who were selected for consultation regarding study feasibility and communications, and 62 captains are "trainers" who are responsible for updating and training captains and first officers with the latest flight techniques. Captains ranged from 37 to 64 years of age, where the average captain was 52 years old and had been an employee of the airline for over 17 years when the study initiated. Captains in the sample flew five flights per month on average, where the captain flying most averaged almost eight flights per month and the captain flying least averaged just over two flights per month.

The resulting dataset consists of 42,012 flights and 110,489 observations of behavior from January 2013 through March 2015 for the captains sampled.[51] We exclude domestic and repositioning flights from our analysis. Among other variables, we observe fuel (kg) onboard the aircraft at four discrete points in time: departure from the outbound gate, takeoff, landing, and arrival at the inbound gate. In addition, we observe fuel passing through each of the aircraft's engines during taxi, which provides a precise measure of fuel burned while on the ground. We also observe flight duration, flight plan variables (i.e. expected fuel use, flight duration, departure

[49] During the study, Rolls Royce (Controls and Data Services) provided monthly data to VAA. We (the academic researchers) almost always received access to the data within two weeks of the start of the month, and feedback reports were compiled and returned to VAA within 24 hours to be postmarked the following day. VAA subsequently provided post-study data (October 2014 through March 2015) for persistence analysis.

[50] The follow-up survey was designed and administered by the academic researchers alone. Again, captains were assured that data from their responses would be used for research purposes only, that their responses would remain anonymous, and that VAA would not be privy to individual-level information provided by survey respondents.

[51] Efficient Taxiing data is physically stored on QAR cards inside the aircraft, which are removed every 2-4 days to pull data. These cards can corrupt or overwrite themselves, and also can reach full memory capacity before being removed. Therefore, data capture for Efficient Taxi is not complete—exactly 37% of flights are missing data for this metric. The reason for the missing data is purely technical and cannot be influenced by captains. We regress an indicator variable of missing Efficient Taxi data on treatment indicators and find no statistically significant relationship at any meaningful level of confidence (individual and joint $p > 0.4$). Consequently, this phenomenon should not affect results beyond reducing the power of estimates.

destination, and arrival destination), and aircraft type. We control for several flight-level variables—e.g., ports of departure and arrival, weather on departure and arrival, whether the aircraft had just received maintenance (e.g., belly wash, engine change), and aircraft type—as well as captain-level time-varying observables such as current contracted work hours and whether the captain had attended the annual Ops Day training.

3. Results

i. Main results

Table 4 and Figures 2a-2c provide a summary description of captains' performance of the fuel-efficient behaviors before and during the experimental period. A preliminary insight is that the pre-experimental behavioral outcomes are balanced across various study groups (see Tables 2 and 3). For instance, roughly 42% of flight observations were characterized by efficient Fuel Load before the experiment started, and attainment within the experimental groups is approximately 41-43%(Table 4, Row 1). Likewise, figures are similar for Efficient Flight (roughly 31%) and Efficient Taxi (roughly 34%). None of the differences across groups are statistically significant at conventional levels.

A second noteworthy insight is the large difference in behaviors before and during the experiment for the control captains, leading to our first formal result:

Result 1. *Captains in the control group change their behavior considerably after they are informed that they are being monitored.*

Preliminary evidence for this result is contained in Column 1 of Table 4. For example, whereas control captains met the Efficient Flight threshold on 31.1% of flights before the experiment, they met the threshold on 47.6% of flights during the experiment ($p<0.01$). Likewise, control captains implemented Efficient Taxi on 50.7% of flights during the experiment compared to 35.2% before the experiment ($p<0.01$). While the results are not economically large for the Fuel Load variable, they again point in the same direction as the other two measures: after the control captains become aware that their actions are being measured, they increase the precision of their fuel load (44.3% versus 42.1% of flight observations; $p<0.05$). Figures 2a-2c

provide a visual summary of this result, and reinforce the substantial difference in captains' behavior once the experiment began.

While these statistics are certainly consistent with Result 1, we have not yet accounted for the data dependencies that arise from each captain's provision of more than one data point. To control for the panel nature of the data set, we estimate a regression model of the form:

$$EfficientBehavior_{it} = \alpha + Exp_{it} \times T_{it}\beta + X_{it}\gamma + \omega_i + e_{it}$$

where $EfficientBehavior_{it}$ equals one if captain *i* performed the fuel-efficient activity on flight *t*, and equals zero otherwise. Exp_{it} indicates the experimental period, T_{it} represents a vector with indicator variables for the three treatments, X_{it} is a vector of control variables, and ω_i is a captain fixed effect. We include all available and relevant flight variables as controls, which include weather (temperature and condition) on departure and arrival, number of engines on the aircraft, airports of departure and arrival, engine washes and changes, and airframe washes. Additionally, we control for captains' contracted flying hours and whether the captain has completed training.[52]

We estimate the above difference-in-difference model specification for each of the fuel-efficient activities using panel data from January 2013 through September 2014, and we treat the first day of the experiment as February 1, 2014, when monitoring of captains begins. Three different empirical approaches yield qualitatively similar results: linear probability model (LPM), probit, and logit. For ease of interpretation, we only present the results of the LPM in Table 5. Robust standard errors are clustered at the captain level. As an alternative, we present Newey-West standard errors for the same model.

We first note the coefficient estimate of the experimental period ("Expt"), which provides a measure of how the control group changed behavior over time. We find a staggering effect: the control group increased their implementation of Efficient Flight by 14.4 percentage points (46.3% effect, 0.31 standard deviations (σ), $p<0.05$) and of Efficient Taxi by 12.5 percentage points (36% effect, 0.26σ, $p<0.05$). Figures 3a-3c demonstrate the pre-experimental trends (from January 2013 through January 2014) and provide a visual representation of the differences in

[52] There are various types of training courses, foremost of which is time spent in the simulator (majority of training) in which captains must pass assessments; we do not have accurate data on these trainings. We instead control for attendance at the two-day "Ops Day" seminar, a gathering of small groups of pilots (approximately 20 per training) for briefing that includes discussion of the goals and directions of the airline and presentations from various teams, with some informal training for pilots.

implementation of the prescribed metrics before and during the experiment. Across both Fuel Load and Efficient Flight, it is clear that there is no upward trend for any group before the experiment started. For Efficient Taxi, we see a slight upward trend, although there is a large increase in the level of implementation during the experimental period across all groups.[53] It is clear that including this trend changes the estimates slightly, especially for the Hawthorne effect in Efficient Taxi—the metric drops by 8.7 percentage points (see Table A1). The Hawthorne effect for Fuel Load increases by 1.5 percentage points and becomes statistically significant ($p<0.05$). We also analyze different time trends (cubic, polynomial, etc.) and they provide very similar estimates to the linear trend analysis.[54]

The above insights lend evidence in favor of a Hawthorne effect, a result consistent with the importance of social pressure in our theoretical structure.[55] They do not, however, shed light on the effectiveness of the treatments in stimulating fuel-efficient behaviors. Results 2-4 address this central question:

Result 2. *Providing captains with information on previous performance moderately improves their fuel efficiency, particularly with respect to Efficient Taxi.*

Result 3. *The inclusion of personalized targets significantly increases captains' implementation of all three measured behaviors: Fuel Load, Efficient Flight, and Efficient Taxi.*

Result 4. *While captains in the prosocial treatment significantly outperform the control group, adding a charitable component does not induce greater effort than personalized targets.*

Overall, Table 5 shows that the effects for all three behaviors are statistically significantly different from the control group at conventional levels for nearly every behavior-treatment combination, both with clustered and with Newey-West standard errors (with a lag of one period). Preliminary evidence of Result 2 can be found in Table 4 and Figures 2-4, which demonstrate that—despite increased performance in Fuel Load and Efficient Flight—the differences between the information and control groups are slight. Yet, there is a considerable

[53] For robustness, we also estimate the specifications in Table 5 with a linear trend—see Table A1.
[54] These analyses are not evidence for the violation of SUTVA, since we reasonably assume that the Hawthorne effect we observe would be applied equally across all groups and not just one or two groups separately.
[55] Table A2 presents three separate Dickey-Fuller tests of a unit root in the pre-experimental data for the three behaviors. The tests provide insight as to whether an upward trend in the pre-experimental data might explain our sizable Hawthorne effects. We collapse the four study groups and analyze each of the three behaviors for 51 weeks preceding the captains' notification of the experiment. For each of the measured behaviors, we reject the null hypothesis that the data exhibit a unit root and therefore argue that the metrics were stationary prior to January 2014.

change in Efficient Taxi implementation between the information and control groups (58.8% versus 50.7%). The standard difference-in-difference estimates in Table 5 complement the raw data in Table 4, demonstrating that the information treatment induces captains to engage in more efficient taxiing. The coefficient estimate suggests that the percentage of flights for which captains receiving the information treatment turned off at least one engine while taxiing to the gate increased by 8.1 percentage points ($p<0.05$) relative to the improvement identified in the control group.

Alternatively, when considering the behavior of captains who received personalized targets in addition to information on previous performance, we observe consistent treatment effects across all three performance metrics. In Tables 4 and 5 and Figures 2-4, we see rather clearly that the targets treatment moved the metrics for each of the three behaviors in the fuel-saving direction. For instance, captains in the targets treatment increased implementation by 3.7 percentage points for Efficient Flight (i.e. a 7.7% treatment effect, 0.074σ, $p<0.05$). Most striking is the effect of the intervention on the occurrence of Efficient Taxi, which occurred on almost 10 percentage points more flights for those in the targets treatment (19.1% effect, 0.194σ, $p<0.01$).[56]

Since each treatment builds upon the last—e.g., feedback in the targets group builds upon that in the information group by adding personalized exogenous targets, holding everything else constant—we "control" for the contents of previous treatments and are therefore able to make comparisons across treatments as well. As shown in Table 5, the information treatment appears to have a positive effect on the incidence of fuel-efficient behaviors compared to the control group, though motivating captains with personalized targets is more effective than using information alone. For instance, the information treatment only significantly increases the Efficient Taxi behavior while targets also significantly increase Efficient Flight (a more difficult metric to achieve). Furthermore, magnitude and significance of the point estimates are increased for captains who receive targets.

That said, prosocial incentives do not appear to provide substantial additional motivation for behavior change beyond targets. The empirical results across the targets and prosocial treatments in Table 5 and Figures 2-4 are very similar. However, captains in these two groups appear to outperform those who received information alone. To statistically validate this

[56] For robustness, we also include specifications where we control for the quadruplet nature of the randomization (see Table A3).

supposition, we pool all captains that receive personalized targets, i.e. target and prosocial treatment groups, and compare the pooled group to the information treatment in an additional regression. We find that receiving targets significantly increases fuel-efficient behavior for Efficient Flight ($p<0.05$) and Efficient Taxi ($p<0.10$). A similar exercise also confirms that prosocial incentives do not significantly improve behavior beyond targets alone. Thus, while information is an important mechanism in encouraging fuel-efficient behavior change, targets add an additional effect that prosocial incentives do not further augment.

In sum, the experimental treatments provide behavioral structure to our theoretical model. Recall that the effect of information on effort in the model depends on the realized difference between estimated and actual fuel efficiency. Given that the estimates suggest a move toward fuel efficiency among captains in the information group (especially with respect to Efficient Taxi), we argue that captains' *ex ante* beliefs regarding their fuel efficiency are optimistic; therefore, information moderately encourages increased fuel efficiency. Our model suggests that targets set above the baseline performance should (weakly) increase effort. Consistent with this conjecture, we find that targets improve captains' attainment of all three behaviors.

Furthermore, the model predicts that the prosocial treatment should increase effort if a captain's altruism is strictly positive and should not affect his effort otherwise. The performance of captains in this treatment group does not significantly exceed that of the captains in the targets treatment on any dimension. Therefore, we cannot conclude that captains' altruism is strictly positive as measured by our experimental manipulation.

Finally, according to the model, captains should allocate effort disproportionately toward behaviors that require the least effort. We know from interviews with captains and airline personnel that Efficient Taxi is the least effortful behavior of the three that were monitored. Our findings support this notion, as the treatment effect sizes for Efficient Taxi are considerably larger than the treatment effect sizes for both Fuel Load and Efficient Flight for all treatment groups.[57]

[57] Note that we are making positive, not normative, statements. In computing welfare effects, one might be concerned with treatment impacts on flight duration and safety. Since there is no variation in safety outcomes (zero incidents or flight diversions due to issues pertaining to fuel), we cannot address this concern. As for flight duration, the treatments did not influence flight duration apart from the reduction in flight time for the prosocial group. In Chapter 4, I aim to address concerns regarding the welfare of captains themselves through measurement of captains' job satisfaction.

ii. Temporal Effects

Importantly, our data provide the opportunity to move beyond short-run substitution effects and explore treatment effects in the longer run. In this sub-section, we conduct a more nuanced investigation of the treatment effects by exploring their persistence as the experiment progresses.[58] Upon doing so, we find a fifth result:

Result 5. *We do not observe decay effects of treatment within the experimental time frame.*

To examine the treatment effects over the course of the experiment, we plot the month-by-month treatment effects in Figures 4a-4c. The largest effects relative to the baseline appear to be in May for Fuel Load and Efficient Flight and in April for Efficient Taxi. That is, the treatment effects appear to be strongest around the middle of the study (and not immediately after monitoring begins), with no consistent pattern of decay for any of the three behaviors.

Although our theory does not have a dynamic decay prediction, given the experimental results in Gneezy and List (2006), Lee and Rupp (2007), Hennig-Schmidt, Sadrieh, and Rockenbach (2010), and Allcott and Rogers (2014), we expected that our treatment effect might decay through time. Indeed, our results are more consonant with Hossain and List (2012), who report that their incentives maintained their influence over several weeks for Chinese manufacturing workers. What our environment shares with Hossain and List's is the context of a repeated intervention whereas the other studies that find a decay effect are typically set within one-shot work environments or weaker reputational environments. We conjecture that repeated interaction with subjects serves to habituate the incented behaviors, thereby diminishing susceptibility to decay effects. Accordingly, this insight serves to enhance our understanding of the generalizability of the decay insights provided in this literature to date.

Another interesting temporal feature in our data is the ability to test for persistence of the treatment effects after the experiment concludes. Inspection of the post-experiment data yields a sixth result:

Result 6. *Treatment effects attenuate or disappear after the treatment is removed, though Hawthorne effects remain high and even increase with the passage of time.*

[58] Relatedly, we also explored a measure of salience in our experiment, namely that behavior changed in the week following receipt of the message and reverted to the mean thereafter. We do not find such an effect.

Once again we find preliminary evidence for this result in Table 4. For instance, while control captains met the Efficient Flight metric on 31.1% of flights before the experiment and 47.6% of flights during the experiment, they actually increased their attainment to 54.8% of flights in the six-month period following the experiment's end date. Similarly, control captains turned off at least one engine while taxiing for 54.7% of flights after the experiment, compared to 50.7% of flights during the experiment and 35.2% before the experiment. This post-experiment increase is not present for Fuel Load, but the original boost in implementation remains after the experiment ends.

Further evidence of persistence is summarized in Table 6, a difference-in-difference specification comparing pre-study behavior to post-study behavior. We see that the control group captains continue to outperform their pre-experimental attainment with significance across all three fuel-efficient behaviors, and even more astoundingly so. The findings indicate that there are no significant differences between control and treatment for Fuel Load and Efficient Flight. However, we still detect significant increases in terms of Efficient Taxi for the targets ($p<0.05$) and prosocial ($p<0.05$) treatment groups, albeit with attenuated treatment effects. These results indicate that the benefits of receiving consistent feedback on fuel-efficient tasks do not persist once the feedback is removed.

iii. Fuel savings

Given the substantial treatment effects during the experimental period of the study, we report an economically significant fuel and cost savings:

Result 7. *The experimental treatments directly led to 704 tons in fuel savings and $553,000 in cost savings for Virgin Atlantic. These estimates dramatically increase after incorporating the estimated Hawthorne effect.*

To provide support for this result, we present two estimations of fuel saved as a result of the experimental treatments. We are in a unique position to use engineering and data-supported fuel estimates to understand the denoted impact of our interventions on efficiency, and we provide both here given that there are pros and cons to each approach.

First, we apply engineering estimates to assess fuel savings without requiring data on actual fuel usage or statistical power to detect differences in fuel use pre- and post-intervention. However, the engineering estimates do not account for *actual* changes to fuel usage as a result of behavior change. While the data-supported estimates do incorporate actual changes to fuel use as a result of the study, the approach is generally one that requires statistical power to detect significant differences in fuel use. Our experimental design was powered to detect differences in fuel-efficient behaviors, not changes in fuel use. As such, we use coefficients that capture average effects of treatments on fuel use without the statistical power to demonstrate significance. Therefore, we use both engineering estimates and data-driven estimates to provide an approximation of fuel saved and CO_2 emissions abated as a result of the treatment groups.

Engineering estimates. VAA projects an average fuel savings of 250 kg per flight as a result of proper execution of Fuel Load. The 0.7%, 2.1% and 2.5% treatment effects for the information, targets, and prosocial incentives groups (respectively) correspond to an increase in the implementation of Fuel Load by 169 flights (saving 250 kg each flight), equivalent to a savings of 42,250 kg of fuel over an eight-month period. Moreover, VAA estimates that an Efficient Flight uses (at least) 500 kg less fuel than the alternative, on average. The effect sizes for the three groups were 1.7%, 3.7%, and 4.7% (respectively), which translates to 323 additional "efficient" flights over the eight-month period, or 161,500 kg in fuel savings. Finally, VAA estimates average fuel wastage of 9 kg per minute if no engines are shut down while taxiing, and the average treatment effects for the three groups were 8.1%, 9.7%, and 8.9%, respectively. Given an average taxi-in time of 8 minutes in our dataset, we approximate fuel savings per flight to be 72 kg. An additional 853 extra flights having met Efficient Taxi corresponds to a fuel savings of 61,400 kg over the eight-month study period.

Summing these savings, the interventions led to just under 266,000 kg of fuel saved over the course of the study. Combining the industry's standard conversion of 3.1497 kg of CO_2 per kg of fuel burned with the February 2014 IATA global jet fuel price of $786 per 1000 kg, we estimate a cost savings of $209,000 and a CO_2 savings of 838,000 kg (i.e. $31,000 environmental savings using $37/ton of CO_2 at 3% discount rate in 2015; Interagency Working Group on Social Cost of Carbon, 2013). The engineering estimates suggest that targets provide the largest benefits to social and private efficiency. These calculations constitute fuel and cost savings stemming directly from the treatments and do not incorporate the sizable Hawthorne effects, which increase the overall cost savings to $1,079,000 and CO_2 savings to 4,324,000 kg. The

savings associated with the Hawthorne effects come from captains having performed Fuel Load on 233 more flights, Efficient Flight on 1,861 more flights, and Efficient Taxi on 1,616 more flights.

Data-supported estimates. The data allow us to estimate actual fuel savings from changes in captains' behavior. We estimate differences in captains' fuel usage from before the experiment to the experimental period within each group. In essence, we employ an Intent-to-Treat approach and use average treatment effects from this difference-in-difference regression to calculate average fuel savings, which we subsequently aggregate.

For Fuel Load, we measure the deviation of the actual fuel load from the "ideal" fuel load—the latter stemming from the double iteration calculation. We identify the average group-level deviation, which is positive if the captain over-fuels relative to the ideal. We then estimate average fuel savings per group, which entails summing the corresponding average treatment effect with the control group's average fuel savings from the pre-experimental to the experimental period (see Table 8). In doing so, we assume that the Hawthorne effect is constant across groups. On average, captains in the control group decreased fuel use relative to the ideal by 128.1 kg per flight, the information group by 98.5 kg per flight, the targets group by 141.3 kg per flight, and the prosocial group by 159.8 kg per flight.

Similarly, for Efficient Flight, we examine changes in captains' fuel use relative to the "ideal" fuel use, or the anticipated fuel use according to the flight plan (adjusted for updates to Fuel Load). We find that captains in the control, information, targets, and prosocial groups reduced in-flight fuel use by 345.2, 371.9, 451.6, and 419.9 kg per flight, respectively. Finally, for Efficient Taxi, we examine changes to fuel use during taxi-in from pre-intervention to the experimental period. Fuel savings per flight amounted to 0.4, 3.7, and 5.1 kg for the control, information, and targets interventions, while the prosocial group increased fuel use during taxi-in by 5 kg.

As a next step, we take these group-level effects and scale them up by the number of flights per treatment group. Put differently, total savings for a given treatment cell are the sum of the average treatment effect and the average Hawthorne effect multiplied by the number of unique flights during the experimental period flown by captains in that group. Results from this exercise are presented in Table 9. Standard error calculations are based on Newey-West standard

errors (lag=1) in the underlying difference-in-difference specifications. Using the data-driven estimates, our interventions led to roughly 6.83 million kg in fuel savings in aggregate. Of these savings, about 1.57 million kg were saved in both the control and information groups, whereas the targets and prosocial group saved more than 1.8 million kg each. Using the same conversions as above (see Table 7), total savings correspond to cost savings of $5.37 million (equivalent to a reduction of 0.56% of overall fuel costs) and CO_2 savings of 21.5 million kg.

Interestingly, there are quite substantial differences between the engineering and data-driven estimates, especially for those that account for Hawthorne effects. The disparity may be attributable to underestimates of average savings from the three behaviors—especially for the Efficient Flight metric—as well as differences in the nature of the estimations. That is, unlike the engineering estimates, the data-supported estimates do not account for differences in percentages of flights for which a behavior was met. Rather, they estimate overall average fuel use changes in the study itself and apply these changes to all flights. Even if we apply the most conservative fuel savings estimates to the changes in behavior, we find that these interventions, especially the target groups, led to remarkable cost-savings and return on investment for the airline.

We calculate an approximate MAC for such behavioral interventions, which is negative (since abatement is highly profitable in this context). Specifically, the MAC (assuming costless interventions) is simply the price per ton of jet fuel divided by 3.15 tons CO_2 per ton of fuel. Using the February 2014 jet fuel price of $786 per ton, we calculate an average MAC of -$250. In other words, each ton of CO_2 abated yields a private cost savings to the airline of $250. Businesses and policymakers should take note of the potential cost-effectiveness of such behavioral interventions in mitigating prominent global externalities.

iv. Treatment effect heterogeneity by prior attainment

Are the results driven by a broad behavioral shift amongst all captains or a handful of captains adjusting completely? To address this question, we explore within-captain differences in attainment from the pre-experimental period to the intervention period. A first result is that the Hawthorne effect is prevalent across captains, as is apparent in Figures 5a-5c. These figures show the change in average attainment of the three behaviors for each control captain. Almost all captains increase their implementation of fuel-efficient behaviors in the experimental period, albeit to varying degrees. Indeed, we find that a majority of captains improve their performance

relative to the baseline for Fuel Load (60% of captains), Efficient Flight (89% of captains), and Efficient Taxi (82% of captains). Looking at the raw data (i.e. without controls), the standard deviations around the mean changes in these behaviors are quite large (Fuel Load: $\mu = 0.036$, $\sigma = 0.105$; Efficient Flight: $\mu = 0.170$, $\sigma = 0.123$; Efficient Taxi: $\mu = 0.147$, $\sigma = 0.149$).

Turning to the question of whether the treatment effects are uniform across captains, we construct similar charts that net out the mean change in behavior of the control group (see Figures 6a-6c). In other words, we deduct the means reported above from each captains' average difference in implementation between the pre-experimental and experimental periods. For example, a captain who implemented Efficient Taxi on 50% of flights before the experiment and 75% during the experiment experienced a 25% increase in attainment, but the Hawthorne effect confounds this increase; therefore, we subtract 14.7%—the average difference among captains in the control group—from 25%, so that the net "effect" on the captain is a 10.3% increase in implementation of Efficient Taxi. Figure 6 displays such within-subject differences in attainment for each of the three measured behaviors across experimental conditions.

There does not appear to be a consistent pattern for Fuel Load and Efficient Flight indicating predictable heterogeneity of treatment effects according to initial attainment levels. However, for Efficient Taxi, relatively low-achieving captains in all three treatment groups appear to outperform similar captains in the control group. Interestingly, for Fuel Load, there is a tendency for the highest-achieving captains to respond negatively to the experiment, perhaps implying a phenomenon akin to "crowding out" of intrinsic motivation. These results are not significant at conventional levels.

4. Discussion

The next time you sit on an airplane next to a policymaker, ask what is the best way to combat pollution externalities. We have posed this question repeatedly working alongside Congresswomen, Senators, and policymakers across governmental agencies. The stock answer is "raise taxes", "create a cap-and-trade scheme", or "make firms install pollution control devices". Not once have we heard: design incentives for workers to produce more sustainably. In this study, we introduce this approach to combating firm-level pollution externalities.

We showcase this approach by implementing a field experiment in a partnership with Virgin Atlantic Airways. The overarching goal was to improve the fuel efficiency of their captains without compromising safety or service quality. While our workplace setting is complex with myriad competing incentives at play, clear measurement of captains' behavior enables innovative strategies to provide the right set of interventions to improve employee productivity and firm performance. Based on our principal-agent model, we randomize three interventions to understand the impact on employee performance of basic informational feedback, exogenous targets associated with said information, and prosocial incentives associated with the above targets and information. We find that all three interventions are successful at inducing fuel-efficient behaviors, and that provision of exogenous targets is the most cost-effective intervention. We conclude that our inexpensive strategies are both a feasible and a profitable means to induce airline captains to fly aircraft more efficiently.

This research speaks to many fields within economics. For example, in labor economics, how best to incent workers to motivate effort in the workplace has been a principal topic of inquiry for decades. The imperfect relationship between employees' effort and productivity renders firms incapable of rewarding effort with precision (Miller, 1992; Lazear, 1999; Malcomson, 1999; Prendergast, 1999). A burgeoning experimental literature on incentives and workplace initiatives attempts to understand the employee-employer relationship and effective means by which employers may increase effort and productivity (see List and Rasul, 2011; Levitt and Neckermann, 2014). We attempt to advance this literature by understanding the separate impacts of basic information, personalized targets, and prosocial incentives on workplace performance in a high-stakes setting among well salaried, experienced, and unionized employees. Our setting does not comprise information asymmetry or team production externalities (i.e. there is no undetected shirking), and therefore there is potential to align individual self-interest with firm efficiency.

This research also has clear policy implications with respect to cost-effective greenhouse gas abatement. We find that the marginal abatement cost (MAC) estimated from no- to low-cost behavioral interventions is around -$250 (using 2014 prices). To our knowledge, this MAC is the lowest currently estimated in academic or policy circles. Thus, such "low-hanging fruits" provide complements—and in some cases perhaps even alternatives—to more traditional approaches to pollution control. Future research should aim to identify additional behavioral motivators to improve the efficiency of workers as a means to minimize abatement costs while

simultaneously reducing the operation costs of firms in an effort to promote win-win strategies for the economy and the environment.

REFERENCES

Air Transport Action Group (2014). Facts and figures. http://www.atag.org/ facts-and-figures.html. Accessed: July 15, 2015.

Al-Ubaydli, O., S. Andersen, U. Gneezy, and J. A. List (2015). Carrots that look like sticks: Toward an understanding of multitasking incentive schemes. *Southern Economic Journal 81*(3), 538-561.

Al-Ubaydli, O. and J. A. List (2015). Do natural field experiments afford researchers more or less control than laboratory experiments? *American Economic Review: Papers & Proceedings 105*(5), 462-66.

Allcott, H. and T. Rogers (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review 104*(10), 3003-37.

Andreoni, J. (1989). Giving with impure altruism: applications to charity and Ricardian equivalence. *Journal of Political Economy*, 1447-1458.

Andreoni, J. (1990). Impure altruism and donations to public goods: a theory of warm-glow giving. *Economic Journal*, 464-477.

Anik, L., L. B. Aknin, M. I. Norton, E. W. Dunn, and J. Quoidbach (2013). Prosocial bonuses increase employee satisfaction and team performance. *PloS one 8*(9), e75509.

Ashraf, N., O. Bandiera, and B. K. Jack (2014). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics 120*, 1-17.

Ashraf, N., O. Bandiera, and S. S. Lee (2014). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior & Organization 100*, 44-63.

Atkin, D., A. Chaudhry, S. Chaudry, A. K. Khandelwal, and E. Verhoogen (2015). Organizational barriers to technology adoption: Evidence from soccer-ball producers in Pakistan. NBER Working Papers 21417.

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy*, 598-614.

Bandiera, O., I. Barankay, and I. Rasul (2007). Incentives for managers and inequality among workers: Evidence from a firm-level experiment. *Quarterly Journal of Economics 122*(2), 729-773.

Bandiera, O., I. Barankay, and I. Rasul (2009). Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica 77*(4), 1047-1094.

Bandiera, O., I. Barankay, and I. Rasul (2010). Social incentives in the workplace. *Review of Economic Studies 77*(2), 417-458.

Bandiera, O., I. Barankay, and I. Rasul (2011). Field experiments with firms. *Journal of Economic Perspectives*, 63-82.

Bandiera, O., I. Barankay, and I. Rasul (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association 11*(5), 1079-1114.

Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy 82*(6), 1063-1093.

Bellemare, C. and B. Shearer (2009). Gift giving and worker productivity: Evidence from a firm-level experiment. *Games and Economic Behavior 67*(1), 233-244.

Bénabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652-1678.

Blanes i Vidal, J. and M. Nossol (2011). Tournaments without prizes: Evidence from personnel records. *Management Science 57*(10), 1721-1736.

Bloom, N., B. Eifert, A. Mahajan, D. McKenzie, and J. Roberts (2013). Does management matter? Evidence from India. *Quarterly Journal of Economics 128* (1), 1-51.

Bloom, N., R. Lemos, R. Sadun, D. Scur, and J. Van Reenen (2014). The new empirical economics of management. *Journal of the European Economic Association 12* (4), 835-876.

Bloom, N., J. Liang, J. Roberts, and Z. J. Ying (2015). Does working from home work? Evidence from a Chinese experiment. *Quarterly Journal of Economics 130*(1), 165-218.

Bloom, N. and J. Van Reenen (2007). Measuring and explaining management practices across firms and countries. *Quarterly Journal of Economics 122* (4), 1351-1408.

Boly, A. (2011). On the incentive effects of monitoring: Evidence from the lab and the field. *Experimental Economics 14*(2), 241-253.

Borenstein, S. (2011). Why can't U.S. airlines make money? *American Economic Review 101*(3), 233-237.

Borenstein, S. and N. L. Rose (2007). How airline markets work...or do they? Regulatory reform in the airline industry. NBER Working Papers 13452.

Borenstein, S. and M. B. Zimmerman (1988). Market incentives for safe commercial airline operation. *American Economic Review 78*(5), 913-935.

Bradler, C., R. Dur, S. Neckermann, and A. Non (2013). Employee recognition and performance: A field experiment. CESifo Working Paper Series 4164.

Burkhardt, U. and B. Kärcher (2011). Global radiative forcing from contrail cirrus. *Nature Climate Change 1*(1), 54-58.

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences 106*(19), 7745-7750.

Button, K. J. and T. G. Weyman-Jones (1992). Ownership structure, institutional organization and measured X-efficiency. *American Economic Review 82* (2), 439-445.

Chandler, D. and A. Kapelner (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization 90*, 123-133.

Charness, G., R. Cobo-Reyes, and A. Sanchez (2014). The effect of charitable giving on workers' performance: Experimental evidence. ThE Papers 14/06, Department of Economic Theory and Economic History of the University of Granada.

Christen, M., G. Iyer, and D. Soberman (2006). Job satisfaction, job performance, and effort: A reexamination using agency theory. *Journal of Marketing 70* (1), 137-150.

Cohn, A., E. Fehr, and L. Goette (2015). Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science 61*(8), 1777-1794.

Delfgaauw, J., R. Dur, A. Non, and W. Verbeke (2014). Dynamic incentive effects of relative performance pay: A field experiment. *Labour Economics 28*, 1-13.

Delfgaauw, J., R. Dur, A. Non, and W. Verbeke (2015). The effects of prize spread and noise in elimination tournaments: A natural field experiment. *Journal of Labor Economics 33*(3), 521-569.

Delfgaauw, J., R. Dur, J. Sol, and W. Verbeke (2013). Tournament incentives in the field: gender differences in the workplace. *Journal of Labor Economics 31* (2), 305-326.

DellaVigna, S., J. A. List, and U. Malmendier (2012). Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics 127* (1), 1-56.

Englmaier, F. and S. G. Leider (2012). Managerial payoff and gift exchange in the field. CESifo Working Paper: Behavioral Economics 3707.

Enkvist, P.A., T. Nauclér, and J. Rosander (2007). A cost curve for greenhouse gas reduction. *McKinsey Quarterly 1*, 34-45.

Federal Aviation Administration (2015). Aviation emissions, impacts and mitigation: A primer.

Freeman, R. B. and M. M. Kleiner (2005). The last American shoe manufacturers: Decreasing productivity and increasing profits in the shift from piece rates to continuous flow production. *Industrial Relations 44*(2), 307-330.

Gibbs, M., S. Neckermann, and C. Siemroth (2014). A field experiment in motivating employee ideas. IZA Discussion Papers 8096, Institute for the Study of Labor (IZA).

Gneezy, U. and J. A. List (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica 74*(5), 1365-1384.

Grant, A. M. and F. Gino (2010). A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior. *Journal of Personality and Social Psychology 98*(6), 946.

Green, J. (2009). The potential for reducing the impact of aviation on climate. *Technology Analysis & Strategic Management 21*(1), 39-59.

Gubler, T., I. Larkin, and L. Pierce (2013). The dirty laundry of employee award programs: Evidence from the field. Harvard Business School NOM Unit Working Paper.

Hamilton, B. H., J. A. Nickerson, and H. Owan (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy 111*(3), 465-497.

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic Literature*, 1009-1055.

Heath, C., R. P. Larrick, and G. Wu (1999). Goals as reference points. *Cognitive Psychology 38*(1), 79-109.

Hennig-Schmidt, H., A. Sadrieh, and B. Rockenbach (2010). In search of workers' real effort reciprocity: a field and a laboratory experiment. *Journal of the European Economic Association 8*(4), 817-837.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, 74-91.

Holmström, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 24-52.

Hossain, T. and J. A. List (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science 58*(12), 2151-2167.

Hubbard, T. N. (2000). The demand for monitoring technologies: The case of trucking. *Quarterly Journal of Economics 115*(2), 533-560.

Hubbard, T. N. (2003). Information, decisions, and productivity: On-board computers and capacity utilization in trucking. *American Economic Review 93* (4), 1328-1353.

Imas, A. (2014). Working for "warm glow": On the benefits and limits of prosocial incentives. *Journal of Public Economics 114*, 14-18.

Interagency Working Group on Social Cost of Carbon (2013). Technical support document: Technical update of the social cost of carbon for regulatory impact analysis—under executive order 12866. https://www.whitehouse.gov/sites/default/files/omb/assets/inforeg/ technical-update-social-cost-of-carbon-for-regulatory-impact-analysis. pdf. Accessed: November 13, 2015.

International Air Transport Association (2014). Fact sheet: Fuel.

International Civil Aviation Organization (2013). Environmental report 2013: Aviation and climate change. http://www.icao.int/environmental-protection/Pages/ EnvReport13.aspx.

Karlan, D., R. Knight, and C. Udry (2015). Consulting and capital experiments with micro and small tailoring enterprises in Ghana. *Journal of Economic Behavior and Organization 118*, 281-302.

Karlan, D. and M. Valdivia (2011). Teaching entrepreneurship: Impact of business training on microfinance clients and institutions. *Review of Economics and Statistics 93*(2), 510-527.

Kosfeld, M. and S. Neckermann (2011). Getting more work for nothing? Symbolic awards and worker performance. *American Economic Journal: Microeconomics*, 86-99.

Kosfeld, M., S. Neckermann, and X. Yang (2014). Knowing that you matter, matters! The interplay of meaning, monetary incentives, and worker recognition. ZEW Centre for European Economic Research Discussion Paper 14-097.

Köszegi, B. and M. Rabin (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics 121*(4), 1133-1165.

Kube, S., M. A. Maréchal, and C. Puppe (2012). The currency of reciprocity: Gift exchange in the workplace. *American Economic Review*, 1644-1662.

Lazear, E. P. (1999). Personnel economics: Past lessons and future directions. *Journal of Labor Economics 17*(2), 199-236.

Lee, D. and N. G. Rupp (2007). Retracting a gift: How does employee effort respond to wage reductions? *Journal of Labor Economics 25* (4), pp. 725-761.

Lee, D. S., D. W. Fahey, P. M. Forster, P. J. Newton, R. C. Wit, L. L. Lim, B. Owen, and R. Sausen (2009). Aviation and global climate change in the 21st century. *Atmospheric Environment 43*(22), 3520-3537.

Lee, D. S. and A. Mas (2012). Long-run impacts of unions on firms: New evidence from financial markets. *Quarterly Journal of Economics 127* (1), 333-378.

Leibenstein, H. (1966). Allocative efficiency vs. "X-efficiency". *American Economic Review 56*(3), 392-415.

Levitt, S. D. and J. A. List (2011). Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments. *American Economic Journal: Applied Economics 3*(1), 224-38.

Levitt, S. D. and S. Neckermann (2014). What field experiments have and have not taught us about managing workers. *Oxford Review of Economic Policy 30*(4), 639-657.

List, J. A. and I. Rasul (2011). *Field Experiments in Labor Economics*, Volume 4 of *Handbook of Labor Economics*, Chapter 2, pp. 103-228. Elsevier.

Locke, E. A. and G. P. Latham (2006). New directions in goal-setting theory. *Current Directions in Psychological Science 15*(5), 265-268.

Malcomson, J. M. (1999). Individual employment contracts. In O. C. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*. Amsterdam: North Holland. Vol. III.

McKinsey (2009). Pathways to a low-carbon economy: Version 2 of the global greenhouse gas abatement cost curve. McKinsey and Company 192.

Miller, G. J. (1992). *Managerial Dilemmas: The Political Economy of Hierarchy*. Cambridge University Press.

Nagin, D. S., J. B. Rebitzer, S. Sanders, and L. J. Taylor (2002). Monitoring, motivation, and management: The determinants of opportunistic behavior in a field experiment. *American Economic Review 92*(4), 850-873.

Penner, J., D. Lister, D. Griggs, D. Dokken, and M. McFarland (1999). Aviation and the global atmosphere: A special report of IPCC Working Groups I and III. Intergovernmental Panel on Climate Change.

Perelman, M. (2011). Retrospectives: X-efficiency. *Journal of Economic Perspectives 25* (4), 211-222.

Porter, M. E. and C. van der Linde (1995). Toward a new conception of the environment-competitiveness relationship. *Journal of Economic Perspectives 9*(4), 97-118.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature 37*(1), 7-63.

Pugno, M. and S. Depedri (2009). Job performance and job satisfaction: An integrated survey. Department of Economics Working Papers 0904, University of Trento, Italia.

Rockoff, J. E., D. O. Staiger, T. J. Kane, and E. S. Taylor (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review 102*(7), 3184-3213.

Ryerson, M. S., M. Hansen, L. Hao, and M. Seelhorst (2015). Landing on empty: Estimating the benefits from reducing fuel uplift in us civil aviation. *Environmental Research Letters 10*(9), 094002.

Schlenker, W. and W. R. Walker (2016). Airports, air pollution, and contemporaneous health. *Review of Economic Studies 83* (2), 768-809.

Shaw, K. (2009). Insider econometrics: A roadmap with stops along the way. *Labour Economics 16*(6), 607-617.

Shearer, B. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *Review of Economic Studies 71*(2), 513-534.

Singh, V. and S. Sharma (2015). Fuel consumption optimization in air transport: A review, classification, critique, simple meta-analysis, and future research implications. *European Transport Research Review 7*(2), 1-24.

Smith, P. (2013). Cockpit confidential: Everything you need to know about air travel: Questions, answers, and reflections. Sourcebooks.

Stern, N. (2007). *The Economics of Climate Change: The Stern Review*. Cambridge Univ. Press.

Stiglitz, J. E. (1975). Incentives, risk, and information: Notes towards a theory of hierarchy. *The Bell Journal of Economics 6*(2), 552-579.

U.S. Enviromental Protection Agency (2015). Inventory of U.S. greenhouse gas emissions and sinks: 1990-2013. http://www3.epa.gov/climatechange/Downloads/ghgemissions/ US-GHG-Inventory-2015-Main-Text.pdf. Accessed: July 15, 2015.

Wiggins, M. W., D. R. Hunter, D. O'Hare, and M. Martinussen (2012). Characteristics of pilots who report deliberate versus inadvertent visual flight into instrument meteorological conditions. *Safety Science 50*(3), 472-477.

FIGURES

FIGURE 1
GLOBAL DESTINATIONS OF VAA

FIGURE 2A
FUEL LOAD, BY TIME PERIOD

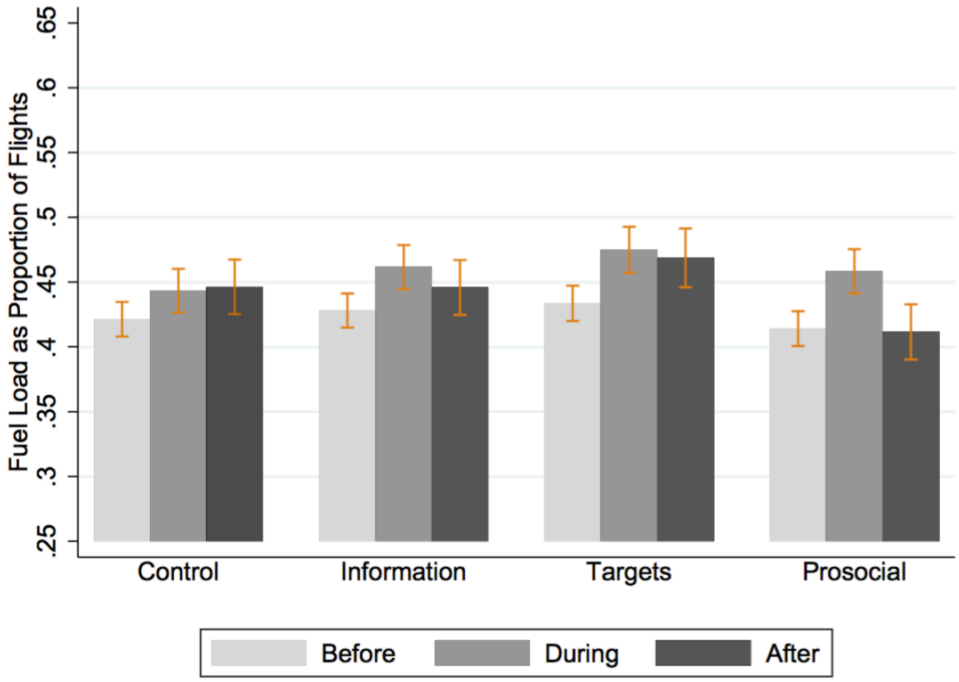


FIGURE 2B
EFFICIENT FLIGHT, BY TIME PERIOD

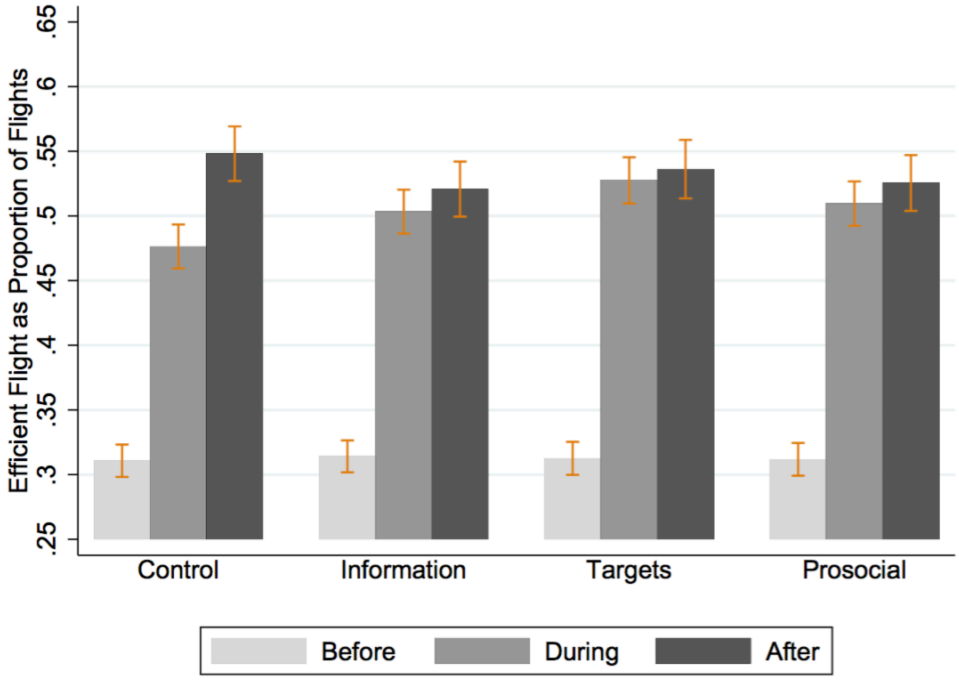


FIGURE 2C
EFFICIENT TAXI, BY TIME PERIOD

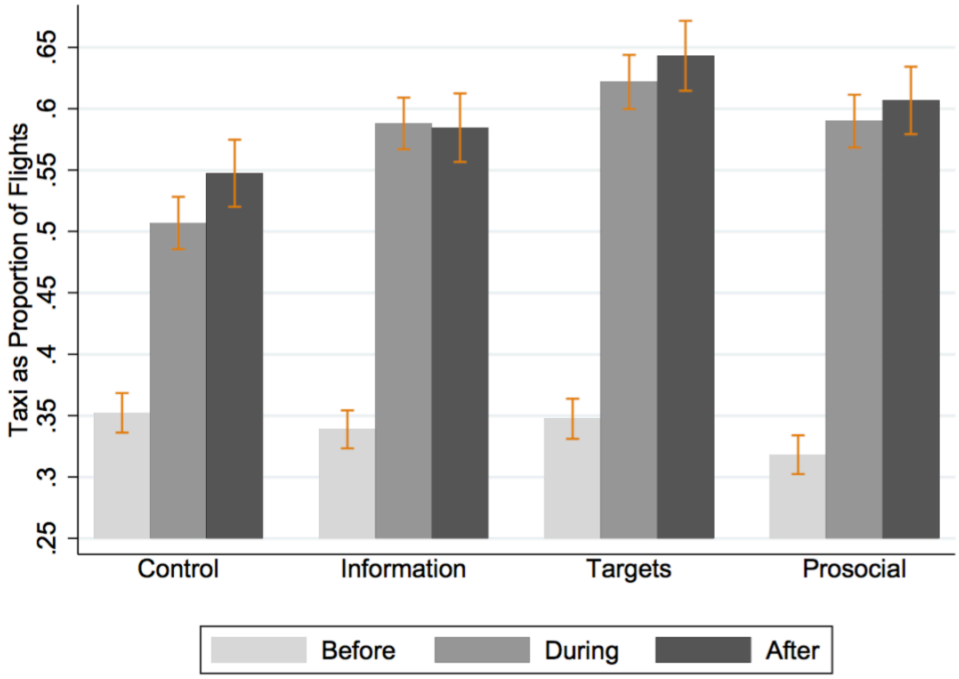


FIGURE 3A
FUEL LOAD BEFORE AND DURING THE EXPERIMENT

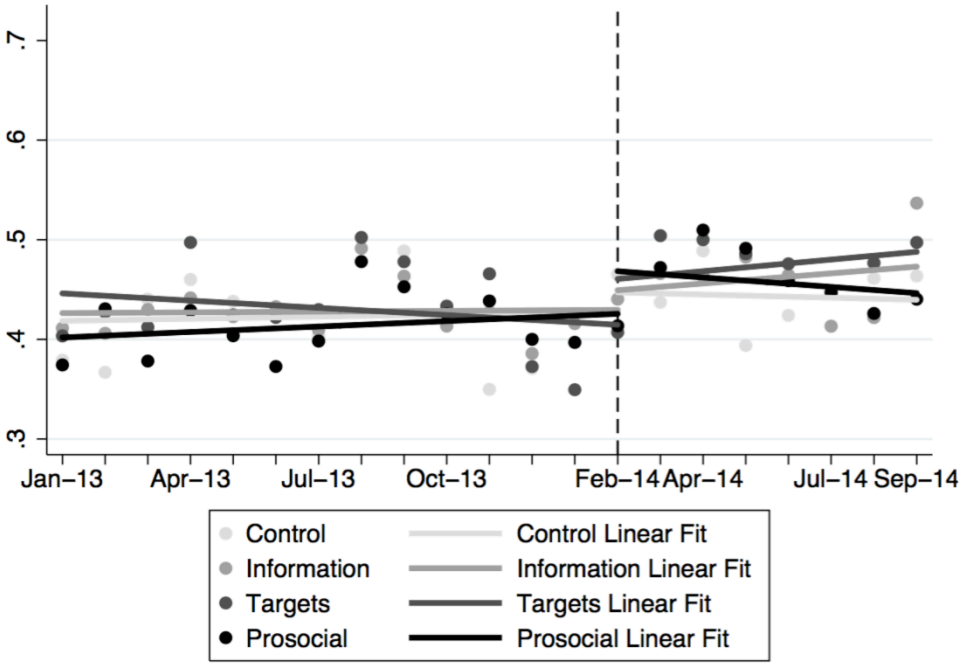


FIGURE 3B
EFFICIENT FLIGHT BEFORE AND DURING THE EXPERIMENT

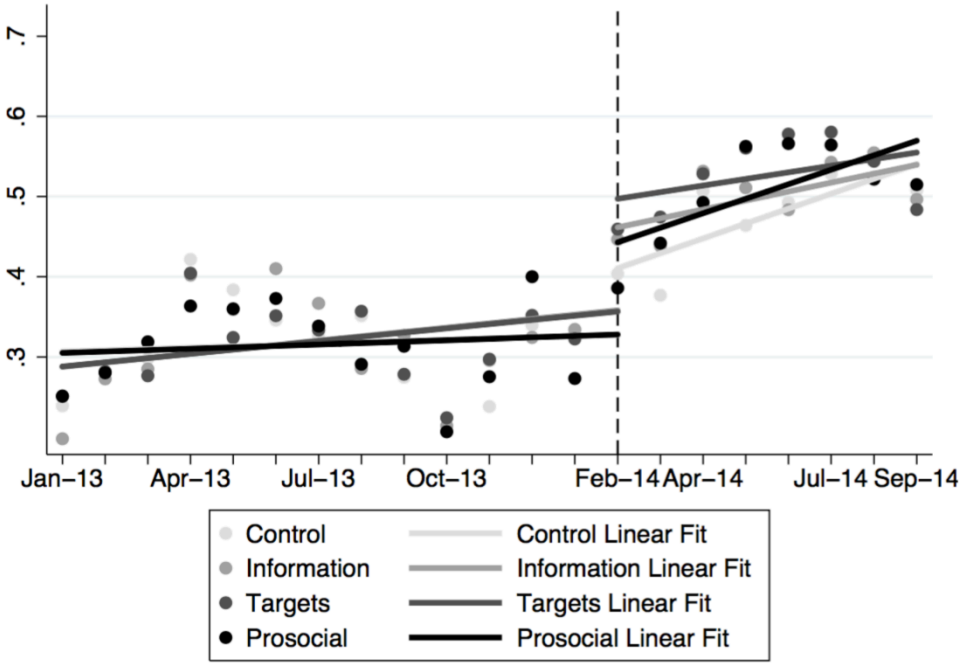


FIGURE 3C
EFFICIENT TAXI BEFORE AND DURING THE EXPERIMENT

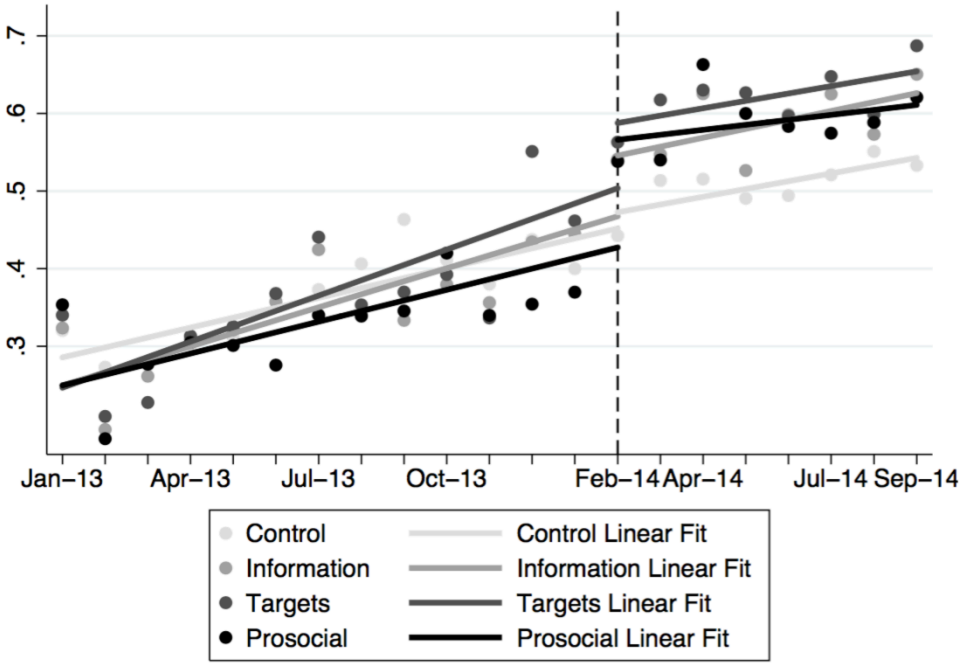


FIGURE 4A
TREATMENT EFFECTS FOR FUEL LOAD DURING THE EXPERIMENT

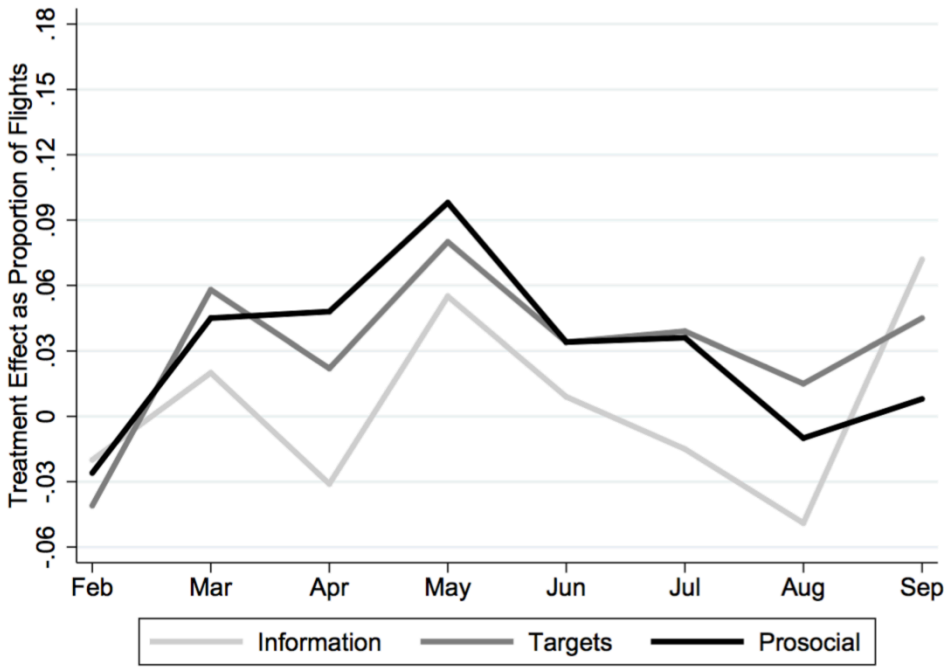


FIGURE 4B
TREATMENT EFFECTS FOR EFFICIENT FLIGHT DURING THE EXPERIMENT

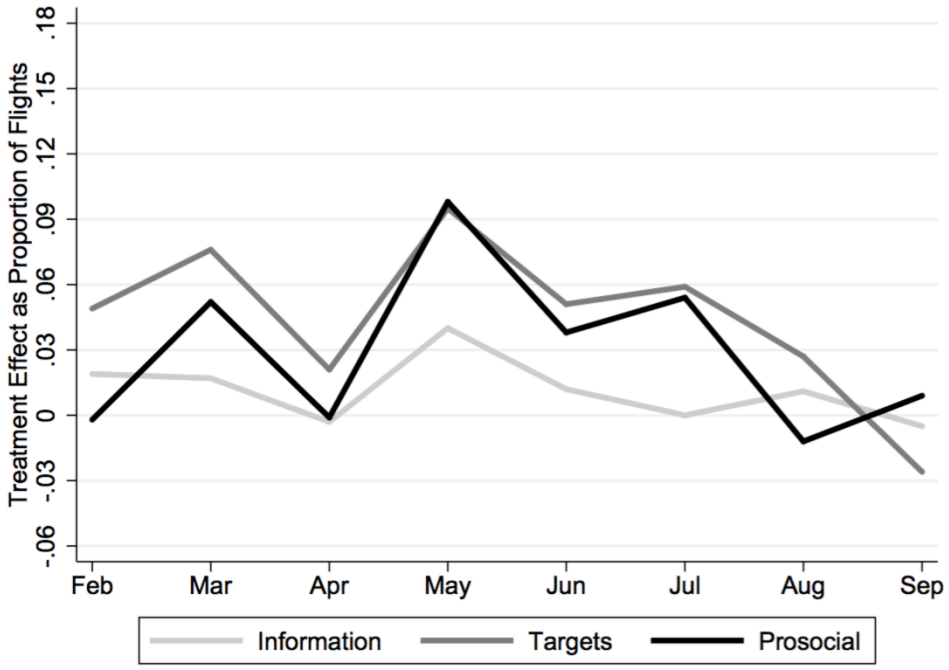


FIGURE 4C
TREATMENT EFFECTS FOR EFFICIENT TAXI DURING THE EXPERIMENT

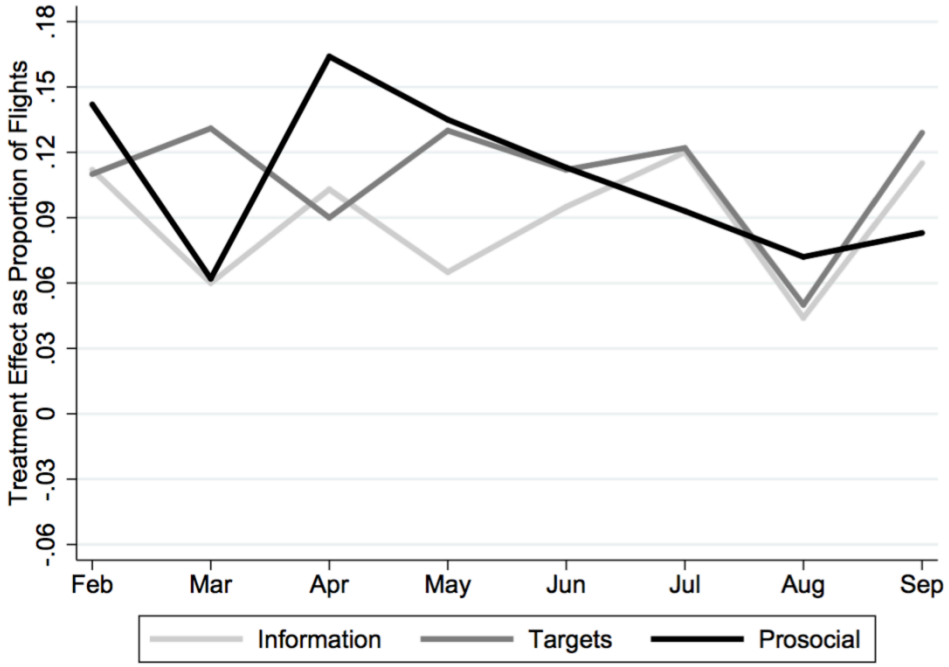


FIGURE 5A
WITHIN-SUBJECT CHANGES IN CONTROL GROUP:
AVERAGE FUEL LOAD IMPLEMENTATION FROM BEFORE TO DURING EXPERIMENT

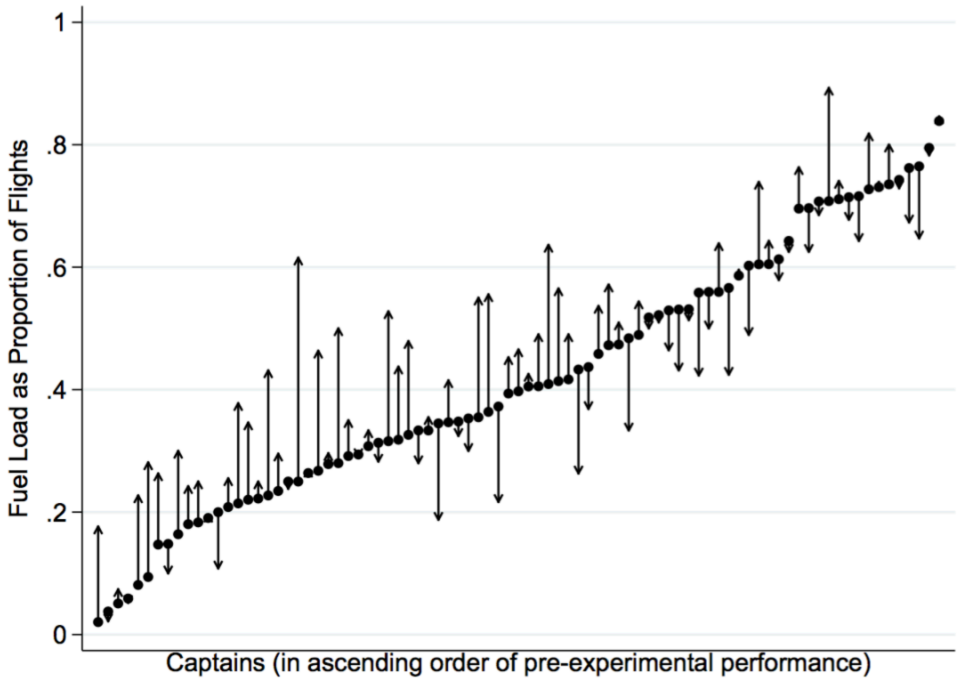


FIGURE 5B
WITHIN-SUBJECT CHANGES IN CONTROL GROUP:
AVERAGE EFFICIENT FLIGHT IMPLEMENTATION FROM BEFORE TO DURING EXPERIMENT

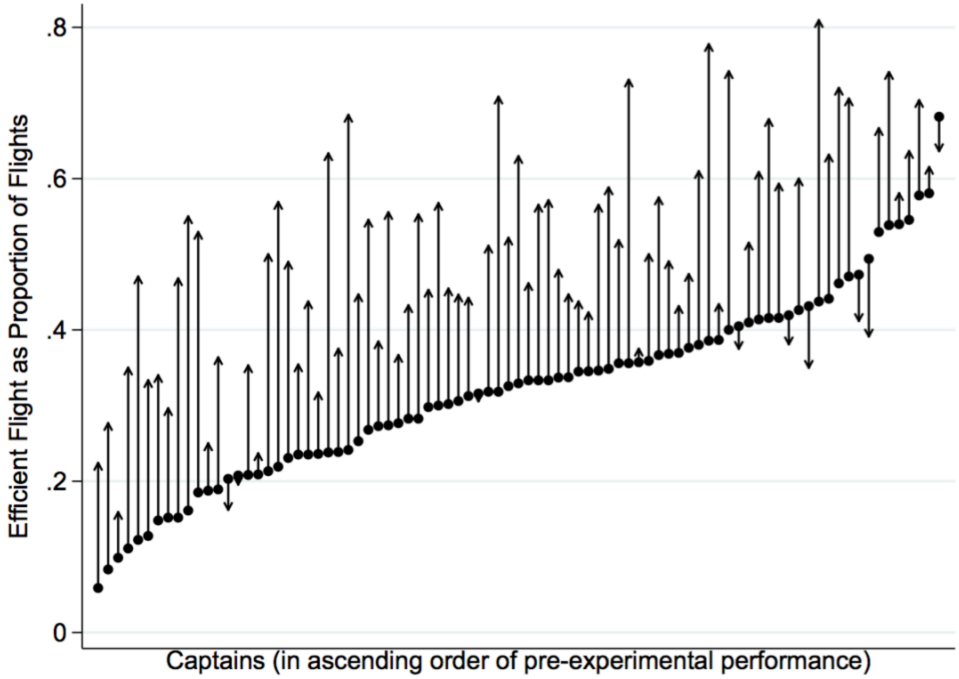


FIGURE 5C
WITHIN-SUBJECT CHANGES IN CONTROL GROUP:
AVERAGE EFFICIENT TAXI IMPLEMENTATION FROM BEFORE TO DURING EXPERIMENT

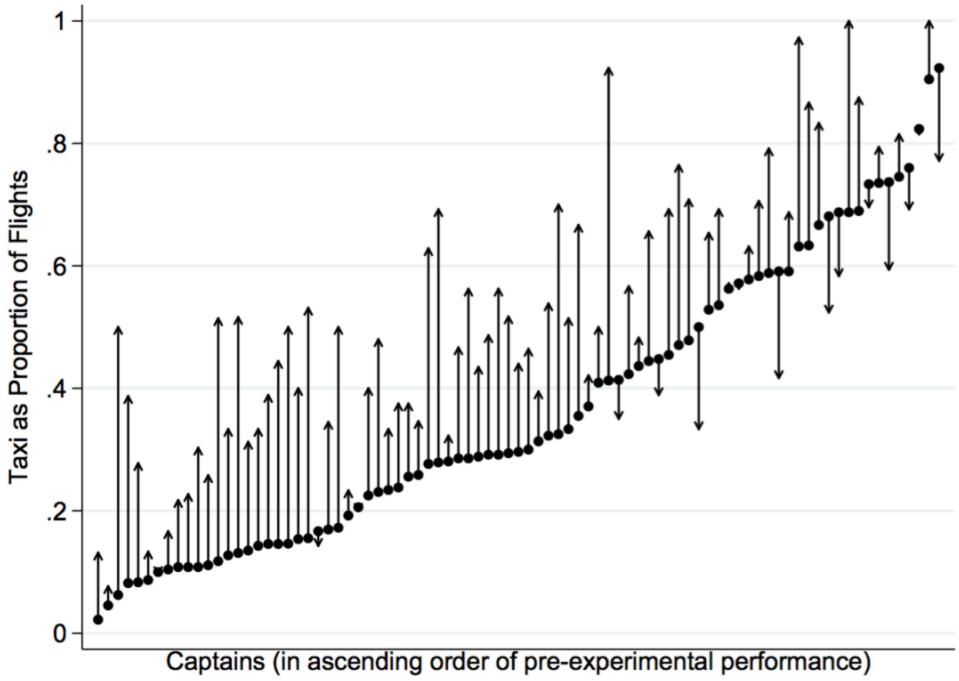


FIGURE 6A
WITHIN-SUBJECT CHANGES IN ALL GROUPS: AVERAGE FUEL LOAD IMPLEMENTATION FROM BEFORE TO DURING THE EXPERIMENT (NET OF RAW HAWTHORNE EFFECTS)

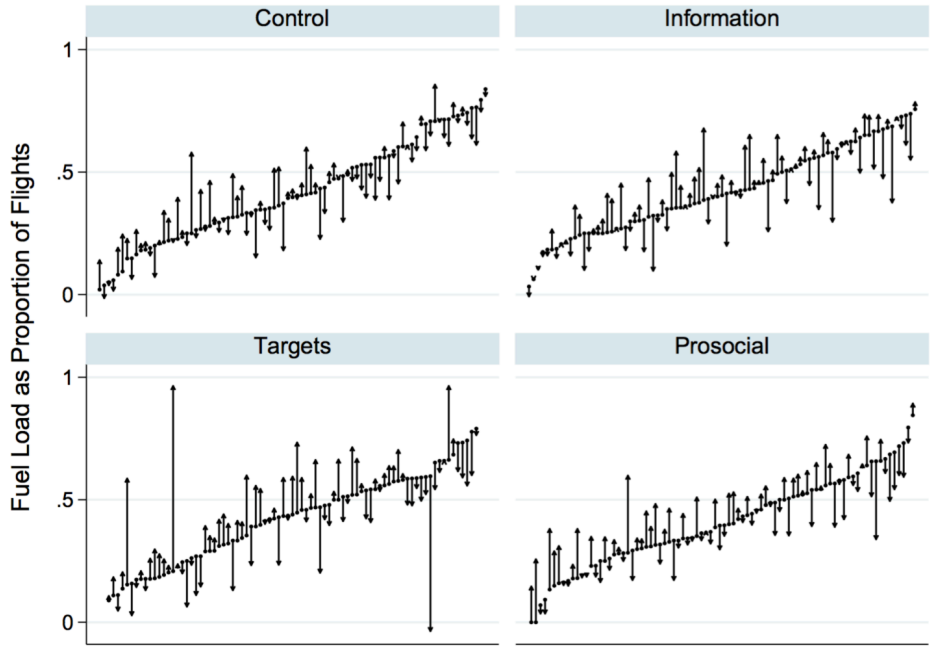


FIGURE 6B

WITHIN-SUBJECT CHANGES IN ALL GROUPS: AVERAGE EFFICIENT FLIGHT IMPLEMENTATION FROM BEFORE TO DURING THE EXPERIMENT (NET OF RAW HAWTHORNE EFFECTS)

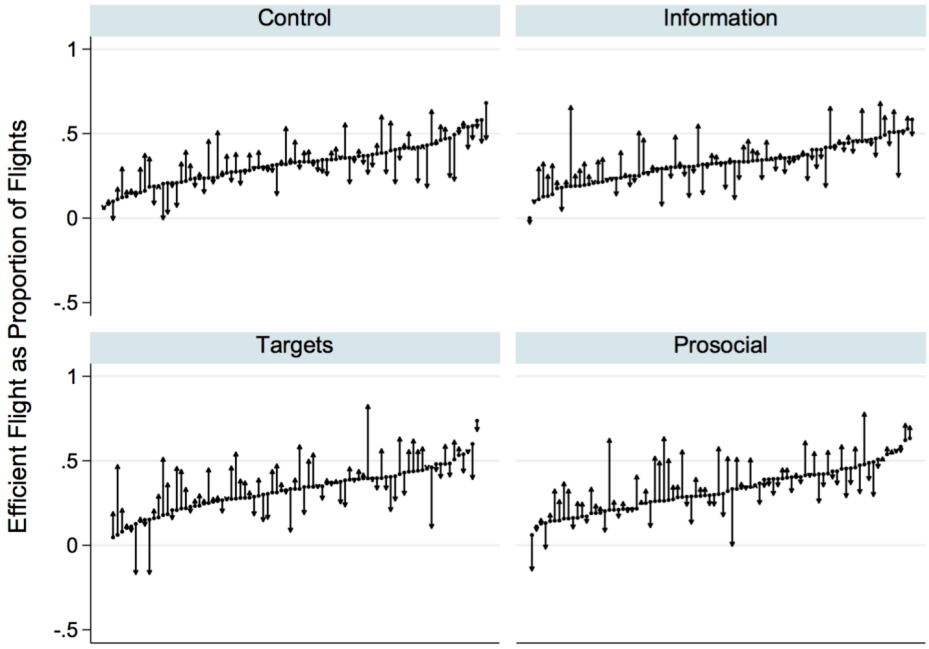
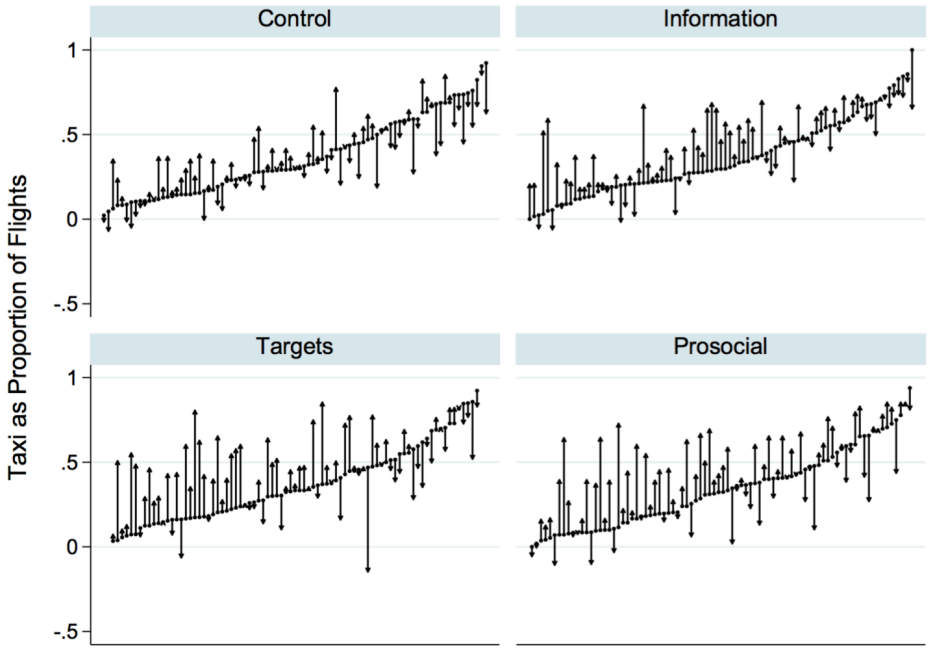


FIGURE 6C

WITHIN-SUBJECT CHANGES IN ALL GROUPS: AVERAGE EFFICIENT TAXI IMPLEMENTATION FROM BEFORE TO DURING THE EXPERIMENT (NET OF RAW HAWTHORNE EFFECTS)



TABLES

TABLE 1
TREATMENT GROUP DESIGN

	Information	Targets	Prosocial Incentives
Control			
Treatment Group 1	✓		
Treatment Group 2	✓	✓	
Treatment Group 3	✓	✓	✓

TABLE 2
BALANCE ON CAPTAIN CHARACTERISTICS

	C: Control	T1: Information	Test of equality: C=T1	T2: Targets	Test of Equality: C=T2	Test of Equality: T1=T2	T3: Prosocial	Test of Equality: C=T3	Test of Equality: T1=T3	Test of Equality: T2=T3
Seniority	177.93 (94.68)	157.16 (97.38)	p=0.161	174.56 (102.00)	p=0.825	p=0.263	171.87 (97.17)	p=0.682	p=0.327	p=0.863
Age	52.23 (5.34)	51.93 (5.10)	p=0.707	51.20 (5.73)	p=0.232	p=0.387	52.31 (5.15)	p=0.926	p=0.633	p=0.193
Trainer	0.165 (0.373)	0.188 (0.393)	p=0.687	0.185 (0.391)	p=0.728	p=0.960	0.202 (0.404)	p=0.527	p=0.817	p=0.780
Trusted Pilot	0.035 (0.186)	0.047 (0.213)	p=0.700	0.025 (0.156)	p=0.690	p=0.440	0.024 (0.153)	p=0.660	p=0.414	p=0.971
Sample	n=85	n=85		n=81			n=84			

Notes: The table reports means and standard deviations (in parentheses) for captains in the four experimental conditions in the pre-experimental data (January 2013-January 2014), in addition to tests of equality for each pair of groups (t-test for continuous variables, chi-square test for indicator variables). *Seniority* and *age* are continuous variables (t-test), while *trainer* and *trusted pilot* are indicator variables. *Seniority* captures the captain's ranking amongst VAA captains. *Age* is the captain's age in years (in 2014). *Trainer* captures whether the captain trains other captains in the latest flight techniques, and *trusted pilot* indicates whether the captain was included in pre-study focus groups.

TABLE 3

BALANCE ON FLIGHT CHARACTERISTICS

	C: Control	T1: Information	Test of equality: C=T1	T2: Targets	Test of Equality: C=T2	Test of Equality: T1=T2	T3: Prosocial	Test of Equality: C=T3	Test of Equality: T1=T3	Test of Equality: T2=T3
Plan Ramp	76,750 (14,993)	78,559 (15,467)	p=0.440	76,666 (14,815)	p=0.971	p=0.764	76,042 (15,587	p=0.422	p=0.294	p=0.793
Engines	3.439 (0.629)	3.483 (0.615)	p=0.648	3.419 (0.640)	p=0.840	p=0.515	3.392 (0.658)	p=0.633	p=0.354	p=0.786
Flights/Month	5.182 (1.372)	5.150 (1.310)	p=0.877	5.305 (1.406)	p=0.571	p=0.465	5.261 (1.328)	p=0.706	p=0.586	p=0.837
Fuel Load	0.417 (0.208)	0.422 (0.175)	p=0.866	0.424 (0.180)	p=0.831	p=0.769	0.408 (0.185)	p=0.956	p=0.616	p=0.589
Eff Flight	0.322 (0.124)	0.322 (0.114)	p=0.979	0.327 (0.130)	p=0.778	p=0.835	0.326 (0.130)	p=0.789	p=0.849	p=0.942
Eff Taxi	0.365 (0.230)	0.359 (0.229)	p=0.874	0.367 (0.222)	p=0.947	p=0.460	0.339 (0.226)	p=0.821	p=0.561	p=0.418
Sample	n=85	n=85		n=81			n=84			

Notes: The table reports means and standard deviations (in parentheses) for captains in the four experimental conditions in the pre-experimental data (January 2013-January 2014), in addition to tests of equality for each pair of groups (t-test for continuous variables, chi-square test for indicator variables). *Plan ramp* measures the amount of fuel anticipated for the entire flight (including taxi-out and taxi-in) and therefore acts as a proxy for distance flown. *Engines* is the average number of engines on aircraft flown. *Flights/Month* is the average number of flights a captain flew in a given month in the thirteen months leading up to the study. *Fuel Load, Eff Flight,* and *Eff Taxi* represent the proportion of each captain's flights on which each of the three fuel-efficient behaviors targeted by the study were met in the pre-experimental period.

TABLE 4
AVERAGE ATTAINMENT OF FUEL LOAD, EFFICIENT FLIGHT, AND EFFICIENT TAXI IN ALL TIME PERIODS (RAW DATA)

	Control (1)	**Treatment 1: Information** (2)	**Treatment 2: Targets** (3)	**Treatment 3: Prosocial** (4)	**All Captains** (5)
Fuel Load					
Pre-intervention	0.421 (0.494) 5258 obs	0.428 (0.495) 5429 obs	0.434 (0.496) 5070 obs	0.414 (0.493) 5140 obs	0.424 (0.494) 20,897 obs
Intervention period	0.443 (0.497) 3321 obs	0.462 (0.499) 3330 obs	0.475 (0.499) 3016 obs	0.458 (0.498) 3258 obs	0.459 (0.498) 12,925 obs
After intervention	0.446 (0.497) 2,140 obs	0.446 (0.497) 2,120 obs	0.469 (0.499) 1,867 obs	0.412 (0.492) 2,063 obs	0.442 (0.497) 8190 obs
Efficient Flight					
Pre-intervention	0.311 (0.463) 5258 obs	0.314 (0.464) 5429 obs	0.313 (0.464) 5070 obs	0.312 (0.463) 5140 obs	0.312 (0.463) 20,897 obs
Intervention period	0.476 (0.500) 3,321 obs	0.503 (0.500) 3,330 obs	0.528 (0.499) 3,016 obs	0.510 (0.499) 3258 obs	0.504 (0.500) 12,925 obs
After intervention	0.548 (0.498) 2140 obs	0.521 (0.500) 2120 obs	0.536 (0.499) 1867 obs	0.525 (0.499) 2063 obs	0.533 (0.499) 8190 obs
Efficient Taxi					
Pre-intervention	0.352 (0.478) 3380 obs	0.339 (0.473) 3596 obs	0.348 (0.476) 3260 obs	0.318 (0.466) 3341 obs	0.339 (0.473) 13,577 obs
Intervention period	0.507 (0.500) 2117 obs	0.588 (0.492) 2,109 obs	0.622 (0.485) 1864 obs	0.590 (0.492) 2014 obs	0.575 (0.494) 8104 obs
After intervention	0.547 (0.498) 1277 obs	0.585 (0.493) 1201 obs	0.643 (0.479) 1,090 obs	0.607 (0.489) 1,218 obs	0.594 (0.489) 4786 obs

Notes: The table reports the proportion of flights for which captains in a given group performed each of the three selected behaviors. Due to random memory errors, Efficient Taxi data is unavailable for 35.0% of pre-experimental flights and 37.2% of post-experimental flights. This missing data is in no way systematic and therefore does not bias the results, though it moderately reduces the power of the Efficient Taxi estimates. Standard deviations are reported in parentheses, which are followed by the total number of observations (flights) from which the summary statistics are calculated.

TABLE 5

TREATMENT EFFECT IDENTIFICATION USING DIFFERENCE-IN-DIFFERENCE REGRESSION

	Fuel Load	Eff Flight	Eff Taxi	Fuel Load	Eff Flight	Eff Taxi
	(1)	(2)	(3)	(4)	(5)	(6)
Expt	0.018	0.144***	0.125**	0.018	0.144***	0.125**
	(0.012)	(0.012)	(0.017)	(0.011)	(0.011)	(0.013)
Expt × Information	0.007	0.017	0.081***	0.007	0.017	0.081***
	(0.017)	(0.016)	(0.025)	(0.015)	(0.014)	(0.017)
Expt × Targets	0.021	0.037**	0.097***	0.021	0.037**	0.097***
	(0.018)	(0.018)	(0.026)	(0.015)	(0.015)	(0.018)
Expt × Prosocial	0.025	0.047**	0.089***	0.025	0.047**	0.089***
	(0.016)	(0.017)	(0.027)	(0.015)	(0.014)	(0.018)
Observations	33,822	33,822	21,681	33,822	33,822	21,681
N	335	335	335	335	335	335
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Standard Errors:						
Clustered	Yes	Yes	Yes			
Newey-West				Yes	Yes	Yes

Notes: The table shows the results of two difference-in-difference regression specifications with captain fixed effects comparing pre-experiment behavior (January 2013-January 2014) to behavior during the experiment (February 2014-September 2014). The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior was performed, and since predicted values are not constrained between 0 and 1, we do not report a constant and instead focus on treatment effects. As such, the coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest was successfully performed. We provide conventional robust standard errors, which are clustered at the captain level, and Newey-West standard errors (lag=1). Total flight observations are provided. Controls include weather on departure and arrival, number of engines on the aircraft, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed training. ***p<0.01 **p<0.05 *p<0.10

TABLE 6

PERSISTENCE: TREATMENT EFFECT IDENTIFICATION USING DIFFERENCE-IN-DIFFERENCE REGRESSION COMPARING BEFORE EXPERIMENT TO AFTER EXPERIMENT FOR ALL EXPERIMENTAL CONDITIONS

	Fuel Load	Eff Flight	Eff Taxi	Fuel Load	Eff Flight	Eff Taxi
	(1)	(2)	(3)	(4)	(5)	(6)
Expt	0.049**	0.215***	0.152***	0.049**	0.215***	0.152***
	(0.024)	(0.021)	(0.035)	(0.021)	(0.020)	(0.029)
Expt × Information	-0.007	-0.021	0.040	-0.007	-0.021	0.040
	(0.020)	(0.021)	(0.030)	(0.017)	(0.016)	(0.021)
Expt × Targets	0.018	0.003	0.057**	0.018	0.003	0.057**
	(0.020)	(0.022)	(0.026)	(0.018)	(0.017)	(0.022)
Expt × Prosocial	-0.020	0.002	0.058**	-0.020	0.002	0.058**
	(0.021)	(0.020)	(0.025)	(0.017)	(0.016)	(0.021)
Observations	29,087	29,087	18,363	29,087	29,087	18,363
N	335	335	335	335	335	335
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Standard Errors:						
Clustered	Yes	Yes	Yes			
Newey-West				Yes	Yes	Yes

Notes: The table shows the results of two difference-in-difference regression specifications with captain fixed effects comparing pre-experiment behavior (January 2013-January 2014) to post-experiment behavior (October 2014-March 2015). The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior was performed, and since predicted values are not constrained between 0 and 1, we do not report a constant and instead focus on treatment effects. As such, the coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest was successfully performed. We provide conventional robust standard errors, which are clustered at the captain level, and Newey-West standard errors (lag=1). Total flight observations are provided. Controls include weather on departure and arrival, number of engines on the aircraft, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed training. ***p<0.01 **p<0.05 *p<0.10

TABLE 7

COSTS OF FUEL USAGE

	Purchase Cost	CO_2 Cost	Total Cost
Fuel (1 ton)	\$786	\$116.55	\$902.55

Notes: The table shows the cost of fuel usage. We use IATA global jet fuel prices in February 2014 (first month of treatment), the EPA estimate of the social cost of carbon of \$37/ton, and the September 30, 2014 exchange rate (\$1 = £0.6167) for all calculations. One ton (1000kg) of fuel emits about 3.15 tons of CO_2. These values are used for calculations of savings in the text.

TABLE 8

DATA-SUPPORTED ESTIMATES OF AVERAGE FUEL SAVINGS PER FLIGHT (IN KG)

	Fuel Load	Efficient Flight	Efficient Taxi
Control	-128.12***	-345.21***	-0.43
	(24.24)	(30.79)	(3.97)
Information	-98.46***	-371.90***	-3.67
	(22.93)	(28.98)	(3.56)
Targets	-141.26***	-451.57***	-5.07
	(22.07)	(27.69)	(4.14)
Prosocial	-159.77***	-419.94***	5.00
	(24.14)	(29.95)	(3.87)

Notes: The table presents estimates of average fuel savings by treatment group in kilograms. Savings are based on regression coefficients from a difference-in-difference specification with captain fixed effects comparing pre-experiment behavior (January 2013-January 2014) to behavior during the experiment (February 2014-September 2014). The Fuel Load and Efficient Flight dependent variables represent the deviation from ideal fuel, whereas the Efficient Taxi dependent variable is simply the fuel used during taxi in. We calculate fuel savings with an Intent-to-Treat approach where we sum the regression coefficient of each group (i.e. the group's average treatment effect) and the average Hawthorne effect (i.e. the coefficient of the Experimental-period indicator). In other words, we assume that the Hawthorne effect is constant across groups. Standard error calculations are based on Newey-West standard errors (lag=1). Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed training. ***p<0.01 **p<0.05 *p<0.10

TABLE 9

DATA-SUPPORTED ESTIMATES OF TOTAL FUEL SAVINGS (IN TONS)

	Fuel Load	Efficient Flight	Efficient Taxi	Total	Per Flight
Control	-425.48***	-1146.46***	-1.45	-1573.38	-0.474
	(80.51)	(102.26)	(13.17)		
Information	-327.88***	-1238.44	-12.21	-1578.52	-0.474
	(76.37)	(96.51)	(11.86)		
Targets	-426.03***	-1361.95***	-15.28	-1803.26	-0.598
	(66.58)	(83.53)	(12.50)		
Prosocial	-520.53	-1368.16***	16.30	-1872.39	-0.575
	(78.65)	(97.58)	(12.60)		
Total	-1699.91	-5115.01	-12.63	-6827.55	-0.828

Notes: The table presents estimates of total fuel savings by treatment group. Savings are based on regression coefficients from a difference-in-difference specification with captain fixed effects comparing pre-experiment behavior (January 2013-January 2014) to behavior during the experiment (February 2014-September 2014). The dependent variable is the deviation from ideal fuel usage in each of the three flight periods described in the text. We calculate fuel savings with an Intent-to-Treat approach where regression coefficients for each group (i.e. the group's average treatment effect) plus the average Hawthorne effect (i.e. the coefficient of the 'Expt' indicator) are multiplied by the number of flights in each group (3321, 3330, 3016, and 3258, respectively). In other words, we assume that the Hawthorne effect for each group is proportional to the number of flights flown by that group during the study period. Standard error calculations are based on Newey-West standard errors (lag=1). Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed training. ***p<0.01 **p<0.05 *p<0.10

APPENDIX I: ADDITIONAL TABLES

TABLE A1

TREATMENT EFFECT IDENTIFICATION USING DIFFERENCE-IN-DIFFERENCE REGRESSION, WITH LINEAR TREND

	Fuel Load	Efficient Flight	Efficient Taxi
Expt	0.033**	0.132***	0.038**
	(0.013)	(0.013)	(0.016)
Expt × Information	0.007	0.017	0.079***
	(0.015)	(0.014)	(0.017)
Expt × Targets	0.022	0.037**	0.096***
	(0.015)	(0.015)	(0.018)
Expt × Prosocial	0.025*	0.047***	0.088***
	(0.015)	(0.014)	(0.018)
Observations	33,822	33,822	21,681
N	335	335	335
Controls	Yes	Yes	Yes

Notes: The table shows the results of a panel difference-in-difference regression specification with captain fixed effects and Newey-West standard errors (lag=1), controlling for linear trends in the data. The regressions compare pre-experiment behavior (January 2013-January 2014) to behavior during the experiment 'Expt' (February 2014-September 2014). The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior was performed, and since predicted values are not constrained between 0 and 1, we do not report a constant and instead focus on treatment effects. As such, the coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest was successfully performed. Robust errors are clustered at the captain level. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed training. ***p<0.01 **p<0.05 *p<0.10

TABLE A2

TEST OF A UNIT ROOT OF PRE-EXPERIMENTAL BEHAVIORS

	Fuel Load	Efficient Flight	Efficient Taxi
Z(t)	-3.765***	-2.562*	-6.431***
	(0.003)	(0.010)	(0.000)
Observations (wks)	51	51	51

Notes: The table shows the Dickey-Fuller (DF) test for a unit root for the 51 weeks before the experiment started, collapsing all the groups into one for each behavior. The null of the DF test is a unit root, and the rejection of the null is that the data follows a random walk. Z(t) is the DF test statistic. ***p<0.01 **p<0.05 *p<0.10

TABLE A3

TREATMENT EFFECT IDENTIFICATION USING DIFFERENCE-IN-DIFFERENCE REGRESSION, WITH QUADRUPLET CONTROLS

	Fuel Load	Eff Flight	Eff Taxi	Fuel Load	Eff Flight	Eff Taxi
Expt	0.022	0.146***	0.140***	0.022**	0.146***	0.140***
	(0.015)	(0.013)	(0.016)	(0.010)	(0.009)	(0.011)
Expt × Information	0.017	0.023	0.065***	0.017	0.023**	0.065***
	(0.020)	(0.016)	(0.022)	(0.012)	(0.011)	(0.014)
Expt × Targets	0.022	0.041**	0.092***	0.022*	0.041***	0.092***
	(0.020)	(0.016)	(0.022)	(0.012)	(0.012)	(0.015)
Expt × Prosocial	0.011	0.042***	0.078***	0.011	0.042***	0.078***
	(0.019)	(0.016)	(0.023)	(0.012)	(0.011)	(0.014)
Observations	33,822	33,822	21,681	33,822	33,822	21,681
N	335	335	335	335	335	335
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Standard Errors:						
Clustered	Yes	Yes	Yes			
Newey-West				Yes	Yes	Yes

Notes: The table shows the results of two difference-in-difference regression specifications with quadruplet fixed effects comparing pre-experiment behavior (January 2013-January 2014) to behavior during the experiment (February 2014-September 2014). The dependent variables in the regressions are dummies capturing whether the fuel-efficient behavior was performed, and since predicted values are not constrained between 0 and 1, we do not report a constant and instead focus on treatment effects. As such, the coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest was successfully performed. We provide conventional robust standard errors clustered at the captain level and Newey-West standard errors (lag=1). Total flight observations are provided. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed training. ***p<0.01 **p<0.05 *p<0.10

TABLE A4

DIFFERENCE-IN-DIFFERENCE REGRESSION OF FLIGHT TIME

	Flight Time
Expt	1.788***
	(0.522)
Expt × Information	0.011
	(0.687)
Expt × Targets	0.114
	(0.733)
Expt × Prosocial	-1.586*
	(0.849)
Observations	33,822
N	335
Controls	Yes

Notes: The dependent variable in this regression is flight time in minutes. Captain fixed effects are included and Newey-West standard errors (lag = 1) are reported below estimates in parentheses. Total flight observations are provided. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed training. ***p<0.01 **p<0.05 *p<0.10

APPENDIX II: THEORETICAL MODEL

Model Setup

We consider a static-choice problem that determines a captain's chosen effort on the job in a certain period. In our model, we assume that captains—who have vast flying experience—are at an equilibrium fuel usage with respect to their wealth, experience, effort, and concerns for safety, the environment, and company profitability.[59]

A captain faces the following additively separable utility function:

$$U(w,s,e,f,g) = u(w,e,g) + a \cdot v(d(e) \cdot g, g_o, G_{-i}) + y(s,e,f) - c(e) - s(e)$$

where $u(.)$ is utility from monetary wealth, $v(.)$ is utility from giving to charity, $y(.)$ is utility from job performance, $c(.)$ is disutility from exerting effort, and $s(.)$ is disutility from social pressure. Effort is chosen for all three flight tasks, j, i.e. Fuel Load, Efficient Flight, and Efficient Taxi. Captains observe their effort perfectly. They also receive a noisy signal of fuel usage $f_{it} + \epsilon_{it} = \bar{f}_{it}$. f_{it} describes the estimated fuel usage by captain i for flight t, which depends on the chosen effort for the fuel-efficient activities. $\bar{f}_{it}$ is actual fuel use, observed *ex post* by the airline, which also includes a random component.[60] Furthermore, each captain has ideal fuel usage f_I, which is based on his own experience and environmental and firm profit preferences. By revealed preference, the equilibrium pre-study fuel usage is $f_I = \bar{f}$.

Experimental treatments in this study alter three model parameters. First, receiving information on fuel use, $i = 1$ (information), removes the noisiness of the fuel signal, i.e. $f_{it} + (1-i)\epsilon_{it} = \bar{f}_{it}$. Second, provision of a target, $r = 1$ (targets), changes the captain's ideal fuel usage, f_I, because the employer exogenously imposes a target level for attainment. Then, $f_I = f_T$ if $r = 1$, where f_T reflects the signaled optimal usage from the point of view of the airline. Third, in the prosocial behavior treatment, a donation g is made by the airline in the name of the captain. This donation is conditional on meeting that target, which has a probability of

[59] In a survey, captains in the study expressed a concern over fuel usage and fuel cost, both for environmental reasons and company profitability. To become an airline captain requires many years of training and experience within an airline; if a captain loses his job with one airline and seeks employment in another, he loses his prior seniority and must work for many years to reinstate it. Thus, for the sake of their own job security, captains care about minimizing fuel costs.

[60] Due to the vast experience of captains, we assume E(ϵ_{it})=0, i.e. captains predict fuel usage correctly, on average.

$d(e)$ in this treatment.[61] In all other treatments, reaching particular fuel use levels does not lead to donations, i.e. $d(e) = 0$. Parameters and elements of the utility function are explained in more detail below.

(Dis)utility from social pressure. In the spirit of DellaVigna et al. (2012, DLM hereafter) and Bénabou and Tirole (2006), we assume that captains are either affected by social pressure due to their actions being observed or exhibit some sort of social signaling in which they want to appear to be good employees. In this framework, captains are aware of an optimal social effort level, e^{social}. Because exerting effort is costly to the captain and because his actions are imperfectly observed with probability $\pi^{observed} < 1$, generally $e < e^{social}$.[62] In this study, captains in both the control group and the treatment groups are made aware that their actions are monitored and data on their effort are used for an internal academic study. Consequently, we expect the probability of detection of deviations from the social effort level to increase for all participants in the study relative to the pre-study period, i.e. $\pi_{study}^{observed} > \pi_{pre}^{observed}$. We parameterize social pressure as follows:

$$s(e) = \left[\pi^{observed} \cdot \left(e^{social} - e\right) + (1 - \pi^{observed}) \cdot 0\right] \cdot 1\left(e < e^{social}\right)$$

Social pressure decreases utility if the chosen effort level is below the socially optimal effort level of the captain. This disutility is increasing in the distance from the optimal effort level and in the probability of these actions being observed by the airline. The second term is an indicator function implying that unobserved deviations do not lead to disutility. For agents that exert more effort than e^{social}, $s(e)$ simply drops out of their utility function. Consequently, captains can directly impact that level of disutility by exerting more (costly) effort.

Note that $s(e)$ enters the utility of every captain below the social effort level, regardless of treatment assignment. If social pressure is important, even control captains should respond to

[61] Captains can directly influence the probability through their effort. That is, captains can be certain that they do not meet a target if they put in little effort, and they can be certain that they have achieved the target if they put in sufficient effort.
[62] It is plausible to argue that effort is perfectly observed in the aviation industry with modern technology. However, captains likely do not expect these data to be analyzed on a regular basis.

this increased cost of low effort.[63] Because $s(e)$ is orthogonal to treatment, we omit it in the following discussion and in the derivation of comparative statics.[64]

Utility from wealth. Similar to DLM, for wealth w, charitable giving from the airline g for meeting the target (if applicable), and other charitable giving g_o, u is defined as follows:

$$u(w, e, g) := u(w - g_o(d(e) \cdot g) + \tilde{a} \cdot d(e) \cdot g)$$

$$\text{where, } \tilde{a} = \begin{cases} 0 \; if \; a < 0 \\ a \; if \; 0 \leq a \leq 1 \\ 1 \; if \; a > 1 \end{cases}$$

Private consumption is an individual's wealth minus the amount given to charity from that person's wealth (i.e. not from this study). However, to ensure that u is continuously differentiable, we need to account for the effect of charitable donations resulting from our treatments on utility from private consumption. To capture this effect, we multiply the individual's expected donation, $d(e) \cdot g$, by a function of a—a parameter capturing preferences for giving—which we call $\tilde{a}$. As in DLM, the parameter a is non-negative in the case of pure or impure altruism and negative in the case of spite[65], and $\tilde{a}$ is simply a truncated at 0 and 1.

The reasons for creating such boundaries on the term capturing preferences for giving are twofold. First, an individual with spiteful preferences ($a < 0$) does not get less utility from private consumption when he donates to charity than when he does not donate to charity; therefore, $\tilde{a}$ is censored from below at 0. Second, an individual with pure or impure altruistic preferences will get additional utility from his private consumption by giving to charity through our treatment because it corresponds to an outward shift in the budget constraint in the dimension of giving to the chosen charity. However, $\tilde{a}$ is censored from above at 1 because an individual will experience weakly more utility from increases in w than from giving to the chosen charity

[63] Alternatively, we could interpret e^{social} as a level of effort induced by the researcher, leading to experimenter-demand effects. Stated differently, captains in the study could think they are expected to increase effort and not doing so imposes utility costs on them.

[64] Social pressure is additively separable from other utility elements in a linear model. Consequently, it does not affect the sign of comparative statics derived below and, if interactions are present, only attenuates treatment effect estimates.

[65] As defined in Andreoni (1989, 1990), pure and impure altruism capture two possible motivations for giving. The first stems from a preference solely for provision of the public good, so that an individual's donations are entirely crowded out by donations from other sources. Impure altruism, on the other hand, refers to the phenomenon whereby individuals receive direct utility from the act of giving itself, i.e. through "warm glow". Spite, as defined in DLM, exists when an individual gets disutility from donating to the charity.

(i.e. $\frac{\partial u}{\partial W} \geq \frac{\partial u}{\partial g}$). This relation holds since increases in w shift the budget constraint outward in all dimensions—including the charitable giving dimension—so these must be weakly preferred to shifts in only one dimension. This stipulation is important to assume differentiability in u in a standard expected utility framework, as in DLM.

Please note that the amount an individual gives to other charities will be related to the amount that he gives to charity in the context of this study. Captains will smooth their consumption for giving. If a captain normally gives \$100 to charity each year and this year he gives \$10 through the context of the study, we would expect his total giving to be between \$100 and \$110, or $g_o + g \in [100,110]$. The realization of the sum depends on the value of a and whether a stems from pure altruism, impure altruism, or spite. We should expect that an individual who has a negative a value does not donate to charity outside of the context of this study since donating to charity decreases that individual's utility.

Utility from charitable giving. The v term is also adapted from DLM and follows the same properties for each type of individual (pure or impure altruistic and spiteful). The main difference between the v term in this study and that in DLM is that in this study, not everyone has the opportunity to donate to charity (i.e. $d(e) > 0$ for only one treatment group). We also assume that v is separable in its parameters, as follows:

$$v(d(e) \cdot g, g_o, G_{-i}) = v_1(d(e) \cdot g, G_{-i}) + v_2(\theta g_o, G_{-i})$$

where θ is the cost of giving through channels other than the study and G_{-i} is total giving by other individuals. In this specification, v_1 represents utility from giving in the study context and v_2 represents utility from giving from one's personal wealth. Note that $v_1(0, G_{-i}) = 0$ since if $d(e) = 0$, then a captain is not able to donate to the charity through the context of the study, so v_1 should not affect the utility function (similar to the spite case). Based on the arguments made above with respect to consumption smoothing, $v|_{d=0} \leq v|_{d=p}$, $0 = v_1|_{d=0} \leq v_1|_{d=p}$, $v_2|_{d=0} \geq v_2|_{d=p}$. That is, utility from giving is at least as high for those captains for whom $d(e) = p$ as it is for those captains for whom $d(e) = 0$, which follows from our assumption that giving in the study context can only decrease giving from one's own wealth or not affect it at all. Finally, since $\frac{\partial p}{\partial e} > 0$, we have $\frac{\partial v}{\partial e} \geq 0$.

In the case of pure altruism, an individual should get the same utility from giving to charity from his personal wealth as from giving to charity through the context of the study, since the benefit to the charity is identical. In this sense, v can be thought to represent the charity's production function. In the case of impure altruism, an individual should also get the same utility from donating to charity through his personal wealth as he does from donating through the context of the study because the amount donated on his behalf is the same. Lastly, in the case of spite, $g_0 = 0$ since giving to charity decreases utility and so those individuals will not give to charity independently of the study. Note, $v(0) = 0$ because if a person does not give to charity privately then his utility from giving to charity privately is 0.

Utility from job performance. Since captains care about fuel efficiency, and since imposing exogenous targets on performance affects a captain's perception of how well he is doing his job, we include a parameter y capturing job performance.[66] We assume y is separable in safety (s) and fuel (f) because changes in fuel as a result of the study do not affect safety levels, as argued in our assumption above. A captain whose performance exceeds his target will achieve higher utility under this parameter than a captain who does not achieve his target. Similarly, a captain will experience less (more) utility the further below (above) the target is his performance. We therefore incorporate job performance into the model as follows:

$$y\,(s,e,f) = y_1(s) + y_2(e,f) = \; y_1(s) + \; y_2\left(-\bar{f}\middle|-f_I\right)$$

where

$$y_2\left(-\bar{f}\middle|-f_I\right) = \; y_{2m}\left(-\bar{f}\right) + \; y_{2n}\left(-\bar{f}\middle|-f_I\right)$$

and

$$y_{2n}\left(-\bar{f}\middle|-f_I\right) = r \cdot \mu(y_{2m}\left(-\bar{f}\right) - \; y_{2m}(-f_I))$$

Here, y_2 is defined as in Köszegi and Rabin (2006; KR hereafter). We denote the components of y_2 "*m*" and "*n*" to mirror the notation in KR. As in KR, *m* represents the "consumption utility" and *n* represents the "gain-loss utility." These terms are separable across dimensions. Finally, μ is the "universal gain-loss function" and has the associate properties outlined in KR. To be clear, we assume that captains who receive exogenous targets perceive these targets as reference points for their own attainment.

[66] Evidence indicates that influencing job performance positively influences job satisfaction (or utility), whether through increased self-esteem or perceived managerial support for autonomous decision making (Christen, Iyer, and Soberman, 2006; Pugno & Depedri, 2009).

Note that captains get utility from using less fuel $\frac{\partial y2}{\partial f} \leq 0$ and, conditional on receiving a reference point, get utility (disutility) from performing above (below) the target, which increases with distance from the target according to μ. We assume μ is linear and $\mu(x) = \eta x$ if $x > 0$ and $\mu(x) = \eta\lambda x$ if $x \leq 0$ for $\eta > 0, \lambda > 1$, in accordance with theories of loss aversion. Moreover, following naturally from our definition of μ, we assume $y(x) = x$. If a captain does not receive a reference point, his utility does not comprise gain-loss utility, so for these individuals $y_2 = y_{2m}$. That is, if $r = 0$, captains do not receive information regarding ideal performance with respect to fuel efficiency, so their job performance parameter depends solely on fuel consumption.[67]

Additionally, based on industry standards and emphasis on safety—as well as the design of the treatments—we assume that captains' job performance utility from flying safely is constant across treatments, therefore:

$$\frac{\partial y}{\partial s} = S \geq 0$$

(Dis)utility from effort. Finally, *c* represents the cost of effort. Importantly, the individual cost functions for each fuel-efficient task are allowed to differ to convey that various tasks have different costs associated with them. The cost structure is a function of the difficulty of the task itself (e.g., it may be easier to turn off one engine after landing than to have an efficient flight for several hours) and resistance due to previous habit formation (e.g., captains who for many years have not properly performed the Zero Fuel Weight calculation may find it difficult or bothersome to begin doing so). Additionally, the costs for each task are separable since the tasks are done independently. Therefore,

$$c(e) = \sum_j c_j(e_j)$$

For a captain to decrease his fuel use, he must also increase his effort, i.e. $\frac{\partial f}{\partial e} < 0$. Note that $c(e)$ is subtracted in the utility equation, so $\frac{\partial U}{\partial c} < 0$, $\frac{\partial c}{\partial e} > 0$. Based on interviews with captains, the cost of effort increases at an increasing rate. Defining the cost of effort as a

[67] To be clear, given that our reference point is exogenously imposed, one cannot clearly assess whether the individual captain is better off in the targets group than in another group.

quadratic function of effort implies that the cost of effort increases with the amount of effort exerted (i.e. $\frac{\partial^2 c}{\partial e^2} > 0$).

Model Predictions

Captains will choose how much effort to exert based on the treatments (information, targets, prosocial incentives) as in the moral hazard model (see Holmström, 1979). The model is simplified because agents are current employees whose base salaries are not affected by the study. The treatments do affect job satisfaction and charitable giving, however. Different treatments represent different contracts.

We now define $V(-f)$ to be the utility of the firm (the principal) from the perspective of the employee (the agent) as a function of firm costs, i.e. fuel costs. V is highly related to y since an employee's job satisfaction is linked to the well-being of the firm itself. We assume V is independent of treatment status, τ, because the marginal benefit and marginal cost to the firm do not depend directly on treatment, but rather on the amount of fuel used (i.e. for the same level of fuel but two different treatments, V is the same). Additionally, salaries are fixed and donations to charity are paid by an outside donor.

We now define $U(e, \tau)$ to be the utility function under treatment τ with effort e and $\overline{U}$ as a captain's outside option.[68] Let $\dot{e}$ be the pre-study amount of effort and $\ddot{e}$ be the chosen effort under τ. Note that the profit-maximizing principal (VAA) wants to design contracts (treatments) that induce the optimal level of effort from the point of view of the principal. In this case, the principal observes both the outcome (fuel usage) and the effort by the agent, but is restricted from making contractual changes that introduce monetary compensation based on effort levels.

Therefore, the problem becomes:

$$\max_{e, g_0} E[V(-f)]$$

$$s.t.\ E[U(w, s, \ddot{e}, f_I, g, \tau)] \geq \overline{U}$$

[68] Our notation differs slightly from Holmström (1979) since the cost of the action is embedded in the utility function of the agent.

$$and\ \ddot{e}\ \in arg\max_{\ddot{e}'} E[U(w,s,\ddot{e}',f_I,g,\tau)]$$

The first-order condition is $\frac{V\prime(-f)}{U\prime(w,s,\ddot{e},f_I,g,\tau)} = \lambda$ and so $U'(w,s,\ddot{e},f_I,g,\tau) = \frac{V\prime(-f)}{\lambda}$. Captains choose the corresponding effort level that satisfies the marginal conditions.

Proposition 1. *Captains in the control group will change their behavior if they are influenced by social pressure. That is, they will generally increase effort if their effort level is below the social effort level.*

Proof: We argued above that scrutiny due to the intervention is likely to (weakly) increase the probability of detection of sub-optimal effort ($\pi^{observed}$) or the perceived socially optimal effort level (e^{social}), or both. Both effects increase the social cost component of the utility function for captains in all treatment cells, including the control group. Put differently, for a given level of effort $\bar{e} < e^{social}$, the intervention increases the marginal social cost of exerting low effort $\frac{\partial U}{\partial s}|\bar{e}$. Consequently, captains respond to these new marginal conditions and increase their effort if they are below the (perceived) socially optimal level.[69]

Proposition 2. Information will cause captains to increase or decrease their effort and thereby increase or decrease fuel usage (respectively) or choose the outside option, depending on the realization of the difference between estimated (f_{it}) and actual ($\bar{f}_{it}$) fuel usage (i.e. the value of the parameter ϵ_{it}).

Proof: Let the pre-study period be $t = 0$ and the study period be $t = 1$. Assume in period $t = 0$, $\epsilon_{i0} < 0$, then $f_{i0} > \bar{f}_{i0}$, so that when captains receive information in $t = 1$, they learn that $y_{2m}(-\bar{f}) > E[y_{2m}(-\bar{f})]$. In other words, they were more fuel-efficient in $t = 0$ than they had expected to be. Therefore, if they provide the same level of effort in period $t = 1$, they will experience a level of utility greater than their pre-study equilibrium. They pay the same cost of effort but receive more utility from job performance. They will then weakly decrease their chosen level of effort. How much depends on the functional form of the y and c functions and their pre-study effort level. Captains in the information or targets treatments—where wealth and

[69] Because of orthogonality to treatment, the condition of being observed simply increases captains' baseline effort. Furthermore, because utility is additively separable, qualitative findings from the subsequent comparative statics analysis are unchanged. As mentioned above, if there are interactions between social pressure and the treatments, these interactions just attenuate point estimates since all treatments are designed to increase effort against a now greater baseline.

the charities' production functions are independent of effort—will not decrease their effort if y is steeper than c around their chosen values. This scenario is possible since there is a random shock of ϵ_{i0} to their location of $-\bar{f}$ and we are agnostic about the functional form of y. Without the shock, they would not be in equilibrium if y were steeper with respect to effort than c at the chosen level of effort because they could increase effort and pay a slightly higher cost but get much more utility from job satisfaction. They will not choose their outside option since if

$E[U(w, s, \dot{e}, f_I, g, \tau = "pre-study, no\ treatment")] \geq \overline{U}$, then

$E[U(w, s, \dot{e}, f_I, g, \tau = "information")] \geq \overline{U}$.

In other words, they can hold y constant and decrease effort and thereby increase U, while $\overline{U}$ is held fixed.

Now assume $\epsilon_{i0} > 0$, then $f_{i0} < \bar{f}_{i0}$, so when captains receive information, they learn that $y_{2m}(-\bar{f}) < E[y_{2m}(-\bar{f})]$, i.e. they were less fuel-efficient than expected. Therefore, if they provide the same level of effort in period $t = 1$, they will receive below their pre-study equilibrium amount of utility. They pay the same cost of effort but receive less utility from job performance. They will weakly increase their effort if the change in y is more than the change in c, which depends on the functional form of these functions and the captains' pre-study effort levels. They will not increase their effort if c is steeper than y for similar reasons described in the previous case. They will choose their outside option if the change in y leads to $E[U(w, s, \ddot{e}, f_I, g, \tau = "information")] < \overline{U}$, which could occur if increases in effort lead to larger increases in c than in y. Whether it occurs also depends on the outside option.

Finally, assume $\epsilon_{i0} = 0$. Then captains are at their equilibrium with $y_{2m}(-\bar{f}) = y_{2m}(-f_I)$ and do not change their effort.

Proposition 3. *Targets set above pre-study use will cause captains to weakly increase their effort or choose their outside option.*[70]

Proof: Since the target is set above pre-study use (i.e. captains are meeting the targets fewer times than is optimal from the perspective of the firm), upon receiving a target, the captains learn

[70] All targets were set above the pre-study attainment level, so this is the only case we consider.

$f > f_T$ and get reference-dependent loss utility equal to $y_{2n} < 0$. Therefore, captains are strictly below their equilibrium in effort and strictly above in fuel usage since in the pre-study period $y_{2n} = 0$ from the assumption that $f_I = \bar{f}$.

Captains will not increase their effort if the increased cost of effort is larger than the gain from the associated decrease in fuel usage in the job performance function. Captains will increase their effort if the gain from the associated decrease in fuel usage is more than the cost of effort. This depends on the functional form of these parameters, the value of μ, and the captains' initial values during the pre-study period. Their chosen level of effort comes from the first-order condition with $\tau = "receive\ targets."$ Since captains experience a negative utility shock from receiving a target, they will choose the outside option if $E[U(w, s, \ddot{e}, f_I, g, \tau = "receive\ targets")] \leq \overline{U}$.

Proposition 4. *Donations made to charity for meeting targets will weakly increase effort if captains' altruism is strictly positive and the donations do not affect their effort otherwise.*

Proof: Let $V_c(d(e), g)$ be the production function of the charity. Note that in the case of pure altruism $V_c = v_1$, as defined in the previous section. $\forall\ d(e) \cdot g \geq 0, we\ have\ V_c > 0\ and\ V_c = 0$ if and only if $d(e) \cdot g = 0$. Then, captains solve the following optimization problem:

$$\max_{e, g_0} E[V(-f) + \tilde{\alpha} \cdot V_c]$$

$$s.t.\ E[U(w, s, \ddot{e}, f_I, g, \tau)] \geq \overline{U}$$

$$and\ \ddot{e} \in arg\max_{\ddot{e}'} E[U(w, s, \ddot{e}', f_I, g, \tau)]$$

with first-order condition $\frac{V'(-f) + \tilde{\alpha} \cdot V_c{}'}{U\prime(av, y, \tau, e,)} = \lambda$. If a captain has zero altruism, i.e. $\tilde{\alpha} = 0$, then this equation reduces to the original and effort does not increase above the effect described in Proposition 1. If $\tilde{\alpha} > 0$, then the numerator of the first-order condition is weakly larger than the control case. It is strictly larger if $d > 0$. Captains with strictly positive altruism may choose an effort level corresponding to $d = 0$ if the additional cost of increased effort required for meeting the target is more than the gain in utility from donating to charity. The probability of this outcome occurring is decreasing in the level of altruism.

Since λ is a constant, increases in the sum of the production functions of the firm and charity cause increases in effort, $\dot{e} < \ddot{e}$.

Proposition 5. *Captains in the targets and prosocial conditions will choose to increase their effort the most in tasks for which the targets are easiest to meet.*

Proof: Since the firm sets the targets and donations exogenously[71], the utility for meeting a target is constant across tasks. The donation to charity is the same across tasks as exogenously determined, and since the targets are also exogenously determined, the captains believe that the firm values them equally by revealed preference. If the firm did not value them equally, then it would not offer the same reward. However, the cost function is not constant across tasks for reasons described earlier, which implies that the captains will choose to increase their effort on tasks for which targets are easiest to meet.[72] Within the airline context, the least effortful behavior to attain is Efficient Taxi, followed by Fuel Load, then Efficient Flight. The determination of this ordering is based on discussions with airline captains and trusted pilots groups.

[71] Note that the "firm" here refers to both VAA and the academic researchers, who jointly made most decisions with respect to experimental design.

[72] Our theory and interventions are rooted in Holmström's (1979) Informativeness Principle, which states that *any* accessible information about an agent's effort should be used in the design and enforcement of optimal contracts. Our interventions are not aimed at the efficient allocation of effort across these tasks—as proposed in Holmström & Milgrom (1991) and Baker (1992)—since we assume our three behaviors are not substitutable (since they occur during different phases of flight). We acknowledge the possibility that additional fuel-efficient behaviors exist that we do not measure that may be fully or partially neglected due to our treatments.

APPENDIX III: TREATMENT LETTERS

FIGURE A1
TREATMENT GROUP 1: INFORMATION

Fuel and carbon efficiency report for Capt. John Smith

Below is your monthly fuel and carbon efficiency report for **February 2014**

1. ZERO FUEL WEIGHT	2. EFFICIENT FLIGHT	3. REDUCED ENGINE TAXY IN
Proportion of flights for which the ZFW calculation was completed and fuel load adjusted as necessary	*Proportion of flights for which actual fuel use is less than planned fuel use (e.g. optimised speed, altitude etc)*	*Proportion of flights for which at least one engine was shut off during taxy in*
RESULT: **0% of flights**	RESULT: **75% of flights**	RESULT: **25% of flights**

We will continue to keep you updated on your monthly performance for the next **7 months**, John.

Please see reverse side for further details of the three behaviours.

Questions? We are here to help! Please email us at project.uoc@fly.virgin.com.

All data gathered during this study will remain anonymous and confidential. Safety remains the absolute and overriding priority. This study will be carried out within Virgin's existing and highly robust safety standards, using our existing fuel procedures and policies. Captains retain full authority, as they always have done in VAA, to make decisions based on their professional judgment and experience.

FIGURE A2
TREATMENT GROUP 2: TARGETS

Fuel and carbon efficiency report for Capt. John Smith

Below is your monthly fuel and carbon efficiency report for **February 2014**

1. ZERO FUEL WEIGHT

Proportion of flights for which the ZFW calculation was completed and fuel load adjusted as necessary

TARGET: **75% of flights**

RESULT: **0% of flights**

You MISSED your target.

2. EFFICIENT FLIGHT

Proportion of flights for which actual fuel use is less than planned fuel use (e.g. optimised speed, altitude etc)

TARGET: **25% of flights**

RESULT: **75% of flights**

You ACHIEVED your target.

3. REDUCED ENGINE TAXY IN

Proportion of flights for which at least one engine was shut off during taxy in

TARGET: **25% of flights**

RESULT: **25% of flights**

You ACHIEVED your target.

WHAT WAS YOUR OVERALL OUTCOME?

You achieved 2 of your 3 targets last month.

WELL DONE! We will continue to keep you updated on your monthly performance for the next **7 months**, John.

Please continue to fly efficiently next month to achieve your targets.

Please see reverse side for further details of the three behaviours.

Questions? We are here to help! Please email us at project.uoc@fly.virgin.com.

All data gathered during this study will remain anonymous and confidential. Safety remains the absolute and overriding priority. This study will be carried out within Virgin's existing and highly robust safety standards, using our existing fuel procedures and policies. Captains retain full authority, as they always have done in VAA, to make decisions based on their professional judgment and experience.

FIGURE A3
TREATMENT GROUP 3: PROSOCIAL INCENTIVES

Fuel and carbon efficiency report for Capt. John Smith

Below is your monthly fuel and carbon efficiency report for **February 2014**

1. ZERO FUEL WEIGHT

Proportion of flights for which the ZFW calculation was completed and fuel load adjusted as necessary

TARGET: **75% of flights**

RESULT: **0% of flights**

You MISSED your target and missed out on £10 in donations to Charity Name.

2. EFFICIENT FLIGHT

Proportion of flights for which actual fuel use is less than planned fuel use (e.g. optimised speed, altitude etc)

TARGET: **25% of flights**

RESULT: **75% of flights**

You ACHIEVED your target and earned £10 in donations to Charity Name.

3. REDUCED ENGINE TAXY IN

Proportion of flights for which at least one engine was shut off during taxy in

TARGET: **25% of flights**

RESULT: **25% of flights**

You ACHIEVED your target and earned £10 in donations to Charity Name.

WHAT WAS YOUR OVERALL OUTCOME?

Due to your fuel and carbon efficient decision making last month, you achieved 2 of your 3 targets and secured £20 of a possible £30 for your chosen charity, Charity Name.

WELL DONE! For the next **7 months**, you still have the ability to donate **£210 to Charity Name.** Please continue to fly efficiently next month to achieve your targets so your charity does not lose out.

Please see reverse side for further details of the three behaviours.

Questions? We are here to help! Please email us at project.uoc@fly.virgin.com.

All data gathered during this study will remain anonymous and confidential. Safety remains the absolute and overriding priority. This study will be carried out within Virgin's existing and highly robust safety standards, using our existing fuel procedures and policies. Captains retain full authority, as they always have done in VAA, to make decisions based on their professional judgment and experience.

FIGURE A4
REVERSE SIDE OF ALL TREATMENT LETTERS

THE THREE BEHAVIOURS WE ARE MEASURING

Behaviour 1: Zero Fuel Weight Adjustment (ZFW) - Pre Flight

This measure compares Actual Ramp against Plan Ramp adjusted for changes in ZFW. It captures whether a double iteration adjustment has been implemented for ZFW in line with Plan Burn Adjustment and any further amendments to flight plan fuel that have been entered into ACARS. This behaviour has a tolerance of 200kg, which ensures that rounding in the fuel request / loading procedure will not adversely affect the result.

Behaviour 2: Efficient Flight (EF) - During Flight

This measure examines the actual fuel burn per minute compared against the expected fuel burn per minute from OFF to ON (expected fuel burn is Plan Trip adjusted for ZFW). It highlights pilot technique (e.g. optimum settings are realised, optimum levels are sought, speed is optimised, etc.).

Behaviour 3: Reduced Engine Taxy In (RETI) - Post Flight

This measure observes if an engine has been shut down during taxy in. RETI is considered to have taken place if one engine burns less than 70% of the average of other engines during taxy in. If taxy in is shorter than the cool down period required, the flight is omitted, as RETI was not possible.

We hope the above information is beneficial to you. If you require more information about any of the behaviours, please email us at project.uoc@fly.virgin.com.

CHAPTER IV

UNDERSTANDING THE ROLE OF CAPTAINS' PREFERENCES IN THE AVIATION INDUSTRY USING COMPLEMENTARY ARTEFACTUAL AND FRAMED FIELD EXPERIMENTS

By Greer Gosnell

Abstract: We measure risk preferences in a high-stakes professional setting to determine whether firm efficiency is affected by agents' risk preferences. Following an eight-month framed field experiment exploring captains' fuel efficiency, we use a conventional economic measure of risk preferences—in addition to self-reported attitudes toward risk and uncertainty, and captains' social preferences—to explore whether heterogeneity in these measures influences performance, both broadly and with respect to performance feedback and incentives. Captains with higher self-reported risk tolerance are more likely to efficiently load fuel onto the aircraft, contributing to average fuel savings on a per-flight basis. Those with higher levels of altruism respond more strongly to prosocial incentives for the binary incentivized behavior (i.e. on the extensive margin), while altruistic captains receiving non-monetary incentives respond with lower fuel use on the intensive margin. Finally, we infer reference dependence and assess welfare impacts from the interventions through analysis of captains' reported job satisfaction.

Acknowledgments: I am grateful to Glenn Harrison and Robert Metcalfe for their roles in helping me to consider the best methods for eliciting risk preferences. I would again like to thank all parties at Virgin Atlantic Airways—especially Paul Morris and Claire Lambert—as well as support teams at MasterCard and Qualtrics who helped to make the survey possible. Finally, I thank the Templeton Foundation for providing the generous funds to support this research.

1. Introduction

The preferences of economic agents are central to the study of decision making at the microeconomic level. Risk-averse economic agents are assumed to exhibit diminishing marginal utility (of income, consumption goods), and to prefer the safer option when considering two payoffs with equal expected value. While it is a widely accepted tenet in conventional agency theory that agents in the workplace are risk averse—thereby creating problems of moral hazard in contracts with presumably risk-neutral principals[73]—little research exists to identify these agents' risk preferences, nor to understand their effect on a firm's performance outcomes. Risk preferences of employees may be especially important when the job description entails some underlying or explicit element of risk. This paper aims to shed light on the relationships between the preferences of airline captains and their performance in relation to fuel efficiency, both generally and with respect to employer monitoring and interventions. Additionally, we look at the knock-on effects of exogenously induced improvements in performance on captains' utility (i.e. job satisfaction).

The role of risk in the workplace has been empirically investigated in various contexts. For instance, firms respond rationally to increases in the probabilities of both inspection and worker injury on the job (Scholz and Gray, 1990). Workers also respond to risk in the workplace, whether by demanding higher compensation to accept potentially dangerous jobs (e.g., Biddle and Zarkin, 1988; Herzog and Schlottmann, 1990; Smith, Pattanayak, and Van Houtven, 2003; Viscusi and Hersch, 2006) or through broader occupation selection (King, 1974; Shaw, 1996; Ackerberg and Botticini, 2002; Bonin et al., 2007; Grund and Sliwka, 2010; Di Mauro and Musumeci, 2011; Bellemare and Shearer, 2013; Fouarge, Kriechel, and Dohmen, 2014; Falco, 2014). Indeed, several studies have found that public sector employees are more risk averse than private sector employees (Bellante and Link, 1981; Masclet et al., 2009; Buurman et al., 2012). Finally, risk holds ambiguous implications for labor contracts; while some claim it distracts from larger determinants (e.g., Allen and Lueck, 1992, 1995, 1999; Lafontaine and Bhattacharyya, 1995), others claim that, for example, it can play a role in determining contracts if the inputs to productivity are unknown to the principal (Prendergast, 2002).

[73] Agency theory generally posits that principals are risk-neutral while agents are risk averse (Allen and Lueck, 1995), the latter preferring certain income to uncertain income (Stiglitz, 1974). Thus, a moral hazard problem between principal and agent arises due to the cost to principals of risk-averse agents' suboptimal effort levels; that is, the agents' preferences motivate effort levels inconsistent with the first-best outcome, which lead to suboptimal contracts (see e.g., Holmström, 1979; Holmström and Milgrom, 1991).

The risk preferences of workers themselves have been elicited and evaluated in several contexts, primarily in relation to agricultural laborers in the developing world who are generally subject to decision making in risky environments (e.g., Binswanger, 1980; Groom et al., 2008; Tanaka, Camerer, and Nguyen, 2010; Bougherara, Gassmann, and Piet, 2011; Picazo-Tadeo and Wall, 2011; Takahashi, 2013). Outside of this context, far less evidence exists on the implications of heterogeneous risk preferences for outcomes in relevant applied settings, an area of research that is arguably highly relevant and vastly understudied (Jamison et al., 2012).

However, a handful of examples demonstrate a move in this direction through the study of risk preferences and their effects on human capital and employment sorting. Shaw (1996) finds that risk tolerance—proxied by percentage of personal worth invested in risky assets—is associated with inherently risky human capital investment, which leads to income growth. Elston, Harrison, and Rutström (2006) find that 'full-time' entrepreneurs are more risk seeking than 'part-time' entrepreneurs, who are in turn more risk seeking than non-entrepreneurs, a finding that may hold implications for sorting and innovation. Using survey evidence from the German Socioeconomic Panel (SOEP), Cornelissen, Heywood, and Jirjahn (2011) find that self-reported risk tolerance increases worker satisfaction in jobs that pay conditional on performance as opposed to fixed wages. Along the same dimension, Bellemare and Shearer (2010) find that workers in a tree-planting firm—whose daily incomes are subject to high performance-related variability—are relatively risk tolerant compared to the general population.

Fewer studies explore the implications of risk preferences and attitudes for relevant performance measures of professionals in high-skilled settings with inherent risk. For example, Eil and Lien (2014) find that poker players engage in more risky behavior as their wealth strays further from their reference points in either direction, and that they become more conservative in response to being ahead, a finding consistent with reference-dependent loss aversion. Coval and Shumway (2005) find a similar effect amongst proprietary traders, yielding short-term consequences for contract prices. Kumbhakar and Tveterås (2003) find evidence of downside risk aversion among 28 salmon farmers in Norway, leading to significant welfare loss in the industry.

These studies do not elicit measures of risk commonly explored in the economics literature, nor do they directly investigate the effects of risk preferences on subsequent decision

making. The present study bears more similarity to Rustichini et al. (2016), who elicit two types of risk measure among trainee truck drivers and relate these to performance. The authors find that a psychological approach (i.e. assessment of personality traits) outperforms an incentivized economic approach in terms of predicting accidents, with the latter demonstrating no predictive power. Here we investigate a role for incentive-compatible risk preferences and self-reported risk attitudes—in addition to attitudes over uncertainty—in predicting a relevant job performance metric pertaining to fuel efficiency in the aviation sector.[74]

In addition to preferences and attitudes toward risk, some literature has suggested that prosocial motivation may measurably improve job performance. In a study of costly cooperation amongst truck drivers in the United States, Burks et al. (2016) perform an artefactual field experiment[75] in which they observe the second-mover behavior of a sequential prisoner's dilemma game, finding that behavior in the game predicts behavior that aids fellow drivers in the field (though not behavior that serves as a favor to the experimenters). In a developing country context, Dizon-Ross, Dupas, and Robinson (2015) find that altruism (aggregated across individuals at the clinic level)—a metric including decisions from an incentivized dictator game—predicts higher coverage for eligible patients and less leakage to ineligible patients in a bed net distribution program. In this paper, we explore a role for altruism in an industrialized high-skilled labor context where prosocial motivation is directly targeted in a large-scale framed field experiment.

We take a novel approach to the study of employee preferences and attitudes in the workplace, incorporating experimental techniques prominent in the economic risk literature into a professional setting where risk and uncertainty play a prominent role. Specifically, we focus our attention on captains in the commercial airline industry, whose risk-averse decisions may influence the airline's bottom line through excess fuel use. Though there are international and domestic requirements for fuel uplift on an aircraft, captains are ultimately responsible for designating the amount of fuel to be carried onboard a given flight. Fuel loaded above and

[74] This research speaks to the psychology and management literatures on whether personality traits should be considered in personnel selection due to their correlations with various aspects of employee productivity (for a meta-analysis, see Tett, Jackson, and Rothstein, 1991). For example, Ashton (1991) compares the predictive capacity of the oft-cited Big Five traits versus the more narrowly defined Jackson Personality Inventory (JPI), finding that the JPI's 'responsibility' and 'risk-taking' measures are better predictors of self-reported delinquency in entry-level jobs than broader Big Five traits. Tett et al. (1991) argue that personality measures indeed hold an important role for personnel selection and further research should aim to identify which personality measures are relevant for various job types. Here, we explore a role for risk preferences as measured in the economics literature.

[75] For a description of experimental typologies, see Harrison and List (2004).

beyond that required—'contingency fuel'—reflects the downside risks that may induce fuel burn additional to that forecasted in the flight plan (Ryerson et al., 2015).

To investigate the role of risk preferences on fuel efficiency, we administered a survey whose respondents had participated in an eight-month framed field experiment prior to preference elicitation. In the experiment, each captain in the airline was randomly allocated to receive one of four monthly interventions: business as usual (control group), performance feedback, exogenous personalized performance targets, and prosocial incentives to reach said performance targets (for a full description and analysis of this study, see Chapter III).[76]

The subsequent online survey elicited captains' incentive-compatible risk preferences as well as their attitudes toward risk and uncertainty, their altruistic tendencies, and their job satisfaction. We elicited risk preferences using a methodology adapted from Binswanger (1980), Barr and Genicot (2008), and Eckel and Grossman (2008). Respondents are asked to select one gamble from five options that progressively increase in both expected value and payoff variance. We also measure self-reported risk attitudes as in Dohmen et al. (2011), who find that these risk measures correlate with incentive-compatible risk preferences. The self-reported approach has several advantages, such as its negligible impact on project budgets and its ability to capture preferences over longer time horizons. However, there is unfortunately very little evidence comparing the predictive power of these two types of measure on real-world outcomes (Jamison et al., 2012).

We additionally derive measures of uncertainty aversion and altruism. To gauge the former, we use a statement derived directly from the Need for Closure Scale (Webster and Kruglanski, 1994), which is meant to assess the extent to which individuals require predictability and resolution. To assess altruism, we ask subjects to recall the extent of their own charitable donations in 2013 (i.e. the year before the study began) and use the midpoint of the selected range as a proxy for subjects' altruistic preferences.

[76] Our flight-level data capture decision making on how much fuel to carry and whether this exceeds the amount dictated by national and company standards (denoted 'Fuel Load'). In addition, our data capture two additional fuel-relevant behaviors that we perceive to be unrelated to preferences over risk: Efficient Flight (i.e. whether the captain used less fuel during the flight than prescribed by the flight plan, updated for Fuel Load adjustments), and Efficient Taxi (i.e. whether the captain turns off at least one engine while taxiing in after landing). While these behaviors will not be a focal point for the analysis of captains' preferences, they were included as behavioral outcomes in the framed field experiment in Chapter 3 and will therefore be alluded to in our analysis of reference dependence in Section IV.

Finally, we exploit the exogenous variation in reference point provision stemming from the framed field experiment to explore whether captains' preferences are in accordance with reference dependence, a prominent concept first proposed in Prospect Theory (Kahneman and Tversky, 1979). Specifically, we examine whether captains who receive exogenous targets in the study are more satisfied with their jobs if those targets are met. Therefore, we provide commentary on reference dependence outside of the monetary domain and provide supporting evidence of the popular finding to the realm of reference points in job performance.

In line with the findings of Rustichini et al. (2016), our results indicate that risk attitudes predict fuel-loading efficiency better than incentive-compatible risk preferences; while the signs of the coefficients on risk preferences yield modest support for this finding, they are not statistically significant. Uncertainty aversion is a slightly less accurate predictor of fuel efficiency, but still contributes to increased fuel uptake. Additionally, more altruistic captains respond more strongly to prosocial incentives on the extensive margin—that is, they implement the incentivized behavior more frequently—while altruism among captains receiving targets alone reduces fuel uptake on the intensive margin. The latter result provides subtle support for the finding in Ashraf et al. (2014) that altruism may crowd in the effects of non-monetary incentives on social outcomes.

We find some evidence of reference dependence among captains who receive targets: job satisfaction increases with the number of targets met for captains who received targets in the study, while this finding does not hold when considering placebo targets for those who did not. Finally, provision of prosocial incentives appears to have a positive impact on captains' well-being. Since targets and prosocial incentives have very similar impacts on fuel efficiency (see Chapter III), this finding induces a tradeoff between employee satisfaction and intervention costs[77] that airlines considering similar interventions should carefully weigh.

The remainder of the paper is structured as follows. Section II provides motivation for the study of preferences in the context of airline captains and fuel efficiency with a discussion of our priors. Section III describes the methodology for eliciting preferences and attitudes, and section IV presents the study results. Section V briefly discusses study limitations, and section VI concludes with a discussion of results and their implications.

[77] The intervention costs may be viewed as a component in airlines' corporate social responsibility (CSR) strategies to provide further support in favor of implementation of prosocial incentives.

2. How might preferences influence outcomes?

According to principal-agent theory, latent risk preferences—unbeknownst to the principal—may influence an agent's decision frame with respect to productivity-enhancing outcomes. In deciding how much fuel to load onto the plane ('Fuel Load'), for instance, risk or uncertainty aversion may lead airline captains to add excess contingency fuel as a safety reserve in case of events with some well-understood probability distribution (e.g., bad weather) or events whose probabilities are difficult to estimate (e.g., unexpected re-routing or airport closures). Thus, we expect aversion to risk and uncertainty to increase fuel uptake above and beyond the 'ideal' uptake that standard industry calculations would prescribe, leading captains to correctly implement Fuel Load (as defined in Chapter III) on fewer occasions.

In the theory of captains' utility proposed in Chapter III (Appendix II), a captain will respond to prosocial incentives by increasing effort only if he possesses a positive level of altruism, and effort increases with his degree of altruism. Accordingly, we expect captains with higher altruistic preferences to respond more strongly to the prosocial incentives provided to the third treatment group in the framed field experiment, since they have stronger motivation to donate to charity. However, altruistic preferences may influence decision making in more than one direction. For instance, 'impurely altruistic' captains that already donate abundantly to charity from their private budgets may not feel the need to exert costly effort in order to donate through the study, as they have already satisfied their utility from 'warm glow' (Andreoni, 1989). Alternatively, it may be the case that more altruistic captains simply care more about social welfare and therefore perform more strongly on the salient prosocial dimension in spite of incentives, as has been found among nurses and doctors in developing country contexts (Callen et al. 2013; Dizon-Ross et al., 2015; Brock, Lange, and Leonard, 2016). We test these competing hypotheses.

Finally, if a captain's utility were indeed characterized by reference dependence and altruism (in accordance with the theoretical model set forth in Chapter III), we would expect the treatments to influence job satisfaction in two measurable ways. First, we posit that a change in a captain's personal expectations from the status quo to an improved level of attainment can boost performance and, consequently, utility. We therefore expect job satisfaction to increase for those who receive targets and subsequently meet them. In other words, a captain will seek to meet the given targets due to the effect of his own job performance on job satisfaction, and fulfilling those

goals will increase his utility. Second, assuming some positive level of (pure or impure) altruism for all captains, those who donated to charity through the study (i.e. 99% of captains in the experimental group that received prosocial incentives, and no additional captains) are expected to have higher job satisfaction than captains who did not donate money to charity through their performance.

3. Methodology

To gather data on captains' preferences, we administered a post-study debrief survey four months after the final day of monitoring for the framed field experiment. The survey served three main purposes: 1) to elicit perceptions on risk and uncertainty; 2) to gauge subjects' altruism; and 3) to measure subjects' job utility (i.e. satisfaction with their jobs).

i. Background: Risk preference elicitation in economics

There are a number of incentive-compatible risk preference elicitation methods that economists use to identify preference heterogeneity and predictive power, as well as the shapes of utility functions. Perhaps the most commonly used method is the Multiple Price List (MPL), which requires that one assume a functional form for an individual's utility and then estimates latent risk preferences by gathering sufficient data to identify intervals for the assumed curvature parameters. To elicit preferences using this method, the researcher presents subjects with a series of binary choices between lotteries with different expected payoffs, and the subject decides which lottery to play, as in the seminal Holt and Laury (2002) elicitation method. In their version of the MPL, payoffs remain constant across ten rows of binary lottery choices, but the probability of the higher payoff increases, eliciting a 'switch point' that identifies the subjects' degree of risk aversion. While this method is widely used (see Andersen et al., 2006, for a review), several studies have called into question its comprehensibility, especially in developing country contexts (e.g., Jacobsen and Petrie, 2009; Charness and Viceisza, 2016).[78] The Becker-DeGroot-Marschak (1964) method—which elicits an incentive-compatible selling price for a series of lotteries—has been deemed similarly incomprehensible (Harrison and Rutström, 2008).

[78] On the contrary, Ihli et al. (2016) find low inconsistency rates for MPL methods in Uganda.

Another heavily cited method is the random lottery pair (RLP) design. Hey and Orme (1994) implement this design by asking subjects to make a selection (or indicate indifference) for each of 100 pairs of lotteries with fixed prizes of £0, £10, £20, or £30, where the probabilities associated with the outcomes vary. For both the MPL and the RLP methods, one lottery is generally selected for payout once preferences are elicited. This exercise was repeated after a few days, but with the presentation of lotteries ordered differently. Subsequently, they estimated a utility function characterized by expected utility theory to assess its performance non-parametrically using behavioral data. Although comprehensibility is strong with this method, the process is quite lengthy and there is not a straightforward ordinal variable that can be used as an independent variable to assess the role of risk preferences in decision making, as we aim to do here.

We therefore elected to use the ordered lottery selection (OLS) method implemented in Binswanger (1980) and Eckel and Grossman (2008), a method that is both succinct and simple to understand. In this method, the subject is presented with an ordered group of (typically 50-50) gambles that increase both in expected payout and in riskiness (variance), beginning with a certain option where both sides of the coin yield the same payoff. Subjects select one lottery from this ordered set, and their choice implies an interval for the subject's coefficient of relative risk aversion under the assumption of utility functions exhibiting constant relative risk aversion (CRRA). The method is computationally simple and quick to implement since only one decision needs to be made and the familiar 'coin-toss' probabilities remain constant across gambles (Charness, Gneezy, and Imas, 2013). It also provides measures that correlate well with other risk preference elicitation methods (Reynaud and Couture, 2012; Dasgupta et al., 2016) and presents significantly less noise than those elicited using more complex methods (Dave et al., 2010). A potential drawback with the OLS method is that the certain option may serve to anchor expectations, creating a gain-loss frame that may influence individual decision making based on reference dependence (Harrison and Rutström, 2008).

Prior studies that elicit risk preferences as an explanatory variable have used both the MPL and OLS methods.[79] For instance, Rustichini et al. (2016) find that trainee truck drivers' risk preferences elicited using MPL (i.e. switch points) are poor predictors of a number of variables, both economic (credit scores, job persistence) and non-economic (driving accidents,

[79] Here we focus our discussion on incentive-compatible risk preference elicitation methods. Other studies have used hypothetical elicitation methods. Barsky et el. (1997), for example, used hypothetical gambles over lifetime income to construct a risk tolerance measure that they use to predict risky behaviors such as smoking, drinking, not buying insurance, choosing a risky job, and holding risky assets.

smoking). Dohmen et al. (2011) use a variant of the MPL where a single lottery is compared with a 'safe' option that increases for each line, eliciting a switch point that they use to predict self-reported behaviors such as smoking, stockholding, and sports participation. Similar to our study, Cardenas and Carpenter (2013) use the OLS method to extract an ordinal measure of risk for use as a right-hand side variable, demonstrating that baseline risk preferences—devoid of subtleties such as preferences over ambiguity or losses—are uncorrelated with well-being in a developing country context. Here, we use a similar method to the latter in our analysis of risk preferences and fuel efficiency.

ii. Survey measures

Risk preferences. To accomplish our primary experimental objective of eliciting preferences over risk, we asked participants a highly incentivized risk question derived from the Binswanger (1980) method of risk preference elicitation that was later adapted by Eckel and Grossman (2008; "EG" hereafter). We implement the latter version, augmenting the stakes by multiplying each outcome by 2.5 (see Table 1). Increasing the stakes amplifies the incentive to carefully consider one's gamble selection, since higher stakes have been found to increase risk aversion (Holt and Laury, 2002) and the sample considered here likely has a much smaller marginal utility of money than do typical risk experimental samples.[80]

The question asks subjects to select one of five gambles, each with two possible outcomes. Gamble 1 is the most certain (£25 vs. £25), while Gamble 5 is the most risky (£105 vs. -£15), and there is a linear relationship between the expected value of the gamble and its riskiness (as measured by the standard deviation of payoffs). Gambles 4 and 5 hold potential for losses (i.e. the second outcome is negative). Both outcomes are associated with a 50% probability of success, as in EG. The method provides sufficient variance in outcomes to estimate ranges of the coefficient of relative risk aversion *r*, a risk parameter that appears in commonly cited CRRA utility functions.[81] The utility function is as follows:

[80] The average salary of a captain is approximately $175,000-$225,000. This salary range is based on information updated in June 2015: http://www.pilotjobsnetwork.com/jobs/Virgin_Atlantic. Assuming diminishing marginal utility of income, a standard experimental incentive used among student samples is likely to provide far less utility to a high-earning professional.

[81] The ranges for the CRRA coefficient vary slightly from that in Eckel and Grossman (2008) due to differences in the initial endowment relative to the incentives provided across the two studies. Additionally, the method cannot detect *r* values in the risk-seeking range. In our sample, 22 captains selected the riskiest gamble in a follow-up question that placed even less probability on the higher outcome, indicating that there may have been even more risky gambles that would have provided greater utility. The gambles presented cannot capture the extent of such risk-seeking preferences.

$$U(x) = \frac{x^{1-r}}{1-r}$$

where x represents a vector of inputs to the utility function and r represents the coefficient of relative risk aversion, and where r<0 causes U to be convex (risk seeking), r=0 causes U to be linear (risk neutral), and r>0 causes U to be concave (risk averse).[82] The function is characterized by constant relative risk aversion, so that relative risk aversion has the same value regardless of consumption levels; in other words, it captures risk attitudes toward proportional changes in wealth at the current level of wealth. As in Cardenas and Carpenter (2013), we order the gambles from 1 to 5 in order of increasing risk tolerance and use this ordinal variable as our incentive-compatible risk measure in the analysis that follows.

Attitudes toward risk and uncertainty. In addition to the above incentive-compatible risk preference elicitation question, we ask subjects to rate their risk tolerance on an 11-point scale, a question that is taken directly from the German SOEP survey. The question asks subjects to choose a number on a scale from 0 ("unwilling to take risks") to 10 ("fully prepared to take risks") in a manner that reflects their own preferences. Dohmen et al. (2011) demonstrate that this question is a good predictor for decisions in incentive-compatible lottery choice experiments; therefore, we use responses to this question as an alternative measure of risk preferences in our analysis and compare findings across the two measures.

Following the above incentive-compatible and self-reported risk questions, we ask a question to assess uncertainty aversion, which is taken directly from the Need for Closure Scale commonly used by social psychologists (Webster and Kruglanski, 1994). In this question, subjects are asked to rate on an 11-point scale from 0 to 10 how much they agree with the following statement: "*I don't like situations that are uncertain*." We use the subjects' responses to determine whether uncertainty plays a different role to risk in fuel uptake decisions.

Altruism. To assess captains' altruism, we asked two questions regarding individuals' private charitable donations in 2013 and in 2014 outside of the context of the study. Captains indicated their past donation behavior by selecting one of ten multiple-choice intervals ranging from £0 to £200+ (specifically, £0, £1-£10, £11-£20, £21-£30, £31-£40, £41-£50, £51-£75, £76-

[82] If r=1, then U(x)=lnx.

£100, £101-£200, more than £200).[83] We then use the midpoints of these ranges as an approximation for donations in a given year (with £250 as the max). In addition to assessment of heterogeneity in fuel outcomes, the information provided further allows us to ascertain whether the framed field experimental prosocial incentive treatment crowds out private giving.

Job satisfaction. Finally, we asked captains to rank their job satisfaction on a seven-point scale. This question comes directly from the British Household Panel Survey, which allows us to compare the satisfaction of VAA captains to the UK national average. Most importantly, the responses allow us to look at both whether job performance influences job satisfaction—as proposed in the theory in Chapter III—as well as whether the framed field experimental treatments contribute positively or negatively to captains' well-being.

4. Results

i. Data

Following participation in an eight-month framed field experiment (February 2014 through September 2014) testing the effects of informational feedback, personalized exogenous targets, and prosocial incentives for meeting these targets, each of the 335 field experimental participants was prompted to take a study debrief survey. Following termination of the study, all captains had been informed that a 5- to 7-minute follow-up survey would be sent to their company email addresses in early 2015. Each captain received an email with a personalized link to a Qualtrics survey on January 29, 2015, and the survey closed three weeks later.[84]

Out of 335 survey recipients, 202 (60%) answered the job satisfaction questions, 193 (58%) answered the later risk, uncertainty, and altruism questions, and 189 (56%) completed the survey in its entirety. This response rate was achieved after sending each captain up to three emails within four weeks offering incentives as high as £130. Of the survey respondents, 97 (49%) reported having received military training and 102 (51%) reported having received civilian

[83] Since very few individuals in the UK donate more than £100 per year, we created smaller bins for lower donation ranges and larger bins for higher donation ranges, though a majority (55% and 56%) of captains in our sample reported having donated more than £100 in 2013 and 2014, respectively.

[84] The follow-up survey was designed and administered by the academic researchers alone and bore no affiliation to Virgin Atlantic. Captains were assured that data from their responses would be used for academic research purposes only, that their responses would remain anonymous, and that VAA would not be privy to individual-level information provided by survey respondents.

training, and the remaining declined to state; we do not have access to this information for the entire captain population.

Survey sampling did not ensure that respondents' characteristics were independently and identically distributed. In Table 2, we check for balance on time-invariant observables (age, seniority, flying frequency, trainer status, trusted pilot status, and study group) as well as average attainment of the fuel-efficient behavior of interest here (Fuel Load) across survey respondents. We find that captains who participated in the survey are less senior ($p<0.10$ for partial survey completion, $p<0.05$ for full survey completion) and properly implement Fuel Load on 5 percentage points more flights, on average (40.5% for non-participants versus 45.7% for participants who completed the survey, $p<0.01$).

Respondents did not differ across the other observable dimensions, and they are well-balanced across treatment groups in the framed field experiment (see Table 3). Apart from a weak correlation between age and uncertainty aversion (β=0.05, p=0.10), none of the captain observables influence any of our preference measures. The average respondent earned £63.01 and donated £20.85 of their survey earnings to charity.

ii. Captains' risk profiles

We find that incentive-compatible gamble selection is highly predictive of self-reported risk attitudes. In a regression of risk attitudes on gamble selection, an increase of one in a subject's gamble selection leads to an increase of 0.29 points on the 11-point risk attitude scale ($p<0.01$). Thus, our data support the finding of Dohmen et al. (2011), which asserts that incentive-compatible and self-reported risk assessments are correlated. However, we do not see a perfectly linear correlation between gamble selection and risk attitude, as can be seen in Table 4, perhaps signaling variation in loss aversion uncorrelated to risk aversion in our sample.[85]

In the vein of Bellemare and Shearer (2010), we compare the risk preferences of the VAA captain population to those of broader populations in various similar studies. EG find an average gamble choice of 3.45 (σ=1.17, n=256) among the students in their sample, whereas we find that captains are more risk averse with a mean gamble choice of 2.96 (σ=1.48, n=164; $p<0.01$).

[85] In other words, captains who are highly loss averse may avoid Gambles 4 and 5, even if they are highly risk loving. We see a steady increase in self-reported risk attitudes as gamble selection increases from Gamble 1 to Gamble 3, with a decline thereafter, potentially due to some captains' aversion to losses. We do not measure loss aversion in this study.

Given that our sample is mostly male, and that EG find that men are more risk-seeking than females, the difference is even more stark when we compare just the male population of EG (3.76, σ=1.18, n=136) with the male captains in our sample (2.95, σ=1.50, n=160; p<0.01).[86] While it is possible that higher stakes induced more risk-averse responses in the survey, we are also mindful that we are comparing students—who are typically considered low-income—to quite a high-income population, so that the stakes relative to income are likely similar.

Contrarily, the captains' self-reported risk attitudes on the 11-point German SOEP scale are consistent with the findings of Grund and Sliwka (2010), who investigate the correlation between risk attitudes and selection into various employment pay schemes (i.e. the risk-incentive tradeoff). They find that individuals in jobs with pay contingencies based on performance appraisals have higher risk tolerance (4.48, σ=2.47) than those with performance appraisals more generally (4.25, σ=2.44), and the latter are still more risk tolerant than individuals in jobs lacking performance appraisals altogether (3.75, σ=2.48). The captains in our sample report a mean of 3.48 (σ=2.31) on the 0-10 scale of risk tolerance, demonstrating even more risk aversion than the average individual in jobs without performance appraisal in Grund and Sliwka's sample.

iii. Preferences and performance

To examine the effects of a captain's preferences and attitudes toward risk and uncertainty on fuel loading behavior, we regress two dependent variables on our risk and uncertainty parameters using OLS regression (see Table 6). The first dependent variable is Fuel Load, which captures the proportion of a captain's flights for which he properly implemented the zero fuel weight (ZFW) adjustment within a 200 kg margin of error (see Chapter III for further detail). Since the outcome variable is binary, these coefficients represent the output of a linear probability model. The second is the difference between the 'ideal' fuel load (i.e. the exact amount prescribed by the double-iteration ZFW calculation) and the actual fuel load, in kilograms. We look at both incentive-compatible risk preferences and self-reported attitudes toward risk and uncertainty (see Table 5 for a list of variable names and definitions, and see Table 6 for regression outcomes).

While we do not detect statistically significant improvements in Fuel Load implementation (β=0.3 percentage points, p=0.65) or the continuous fuel load measure (β=-22.0

[86] Since there are only 4 females in our sample, we cannot compare across the female populations in the two studies.

kg, p=0.14) based on the incentive-compatible gamble selection, the signs of the coefficients hint that more risk-seeking individuals are more likely to implement Fuel Load and to load less fuel on a given flight. This influence is more firmly supported when we use the self-reported risk attitude as our independent variable. For every increase of one point on the German SOEP risk scale, implementation of Fuel Load increases by 0.9 percentage points (p=0.02) and fuel uptake relative to the ideal decreases by 16.8 kg (p=0.06).[87] Similarly, as uncertainty aversion increases, we see a negative (though not significant) coefficient for Fuel Load (β=-0.3 percentage points, p=0.54) and an increase in the amount of fuel loaded onto the aircraft relative to the ideal (β=14.0 kg, p=0.09). Thus, consistent with our priors, we find evidence that more risk-averse captains are more likely to over-fuel.[88]

In Table 7, we use self-reported donations to charity in 2013 (i.e. the year before the study began) as a proxy for altruism to explore whether subjects' prosocial preferences influence their response to our treatments. In the 13 months prior to the study, more altruistic captains use slightly less fuel, though the effect is not statistically significant (see Columns 1-2 of Table 7). Therefore, altruism does not appear to play a role in fuel efficiency *per se*.

However, we do find that captains vary in their response to treatment based on their revealed altruistic preferences (consistent with Proposition 4 in the theory). We interact the altruism proxy with the difference-in-difference estimator identifying the treatment effects in Chapter III, allowing us to investigate whether altruism influences captains' response to treatment. Once the study begins, we find that captains who receive prosocial incentives for meeting the study targets implement Fuel Load on about 0.1 percentage points more flights per £10 privately donated in 2013 (p=0.01; Table 7, Column 3). That is, these captains improve their Fuel Load implementation on the extensive margin, on which the conditional incentives were

[87] Tett et al. (1991) find that correlations between productivity and personality traits are stronger for military personnel than for non-military personnel. On the contrary, we find that the risk attitudes are more strongly predictive of performance for civilian-trained captains with a coefficient more than double in magnitude; an increase of 1 on the risk attitude scale increases implementation of Fuel Load by 1.5 percentage points (p=0.03) for civilian captains, whereas the corresponding effect for military captains is 0.7 percentage points (p=0.15).

[88] As a robustness check to ensure that the disparities in Fuel Load are in the direction of over- (and not under-) fueling, we regress a dummy variable for whether the captain over-fueled relative to the ideal on incentive-compatible risk preferences, risk attitudes, and uncertainty aversion. An increase of one on the incentive-compatible and self-reported risk scales decreases over-fueling by 0.2 (p=0.49) and 0.5 (p=0.02) percentage points, respectively, while the same increase on the uncertainty aversion scale increases over-fueling by 0.2 percentage points (p=0.41). The magnitudes of these effects increase or remain constant if we allow for the 200 kg margin of error for incentive-compatible risk (β=0.3 percentage points, p=0.67), self-reported risk (β=-0.9 percentage points, p=0.03), and uncertainty aversion (β=0.2 percentage points, p=0.55). Under-fueling outside of the margin of error only occurred on 2.7% of flights during the study period (2.8% for subjects who completed the survey).

predicated.[89] This effect is slightly positive though highly indistinguishable from zero for the information and targets groups. Therefore, we find evidence against a warm-glow 'quota' that may decrease the motivational capacity of prosocial incentives.

Interestingly, however, more altruistic captains who receive targets without receiving prosocial incentives decrease their fuel use relative to the prescribed fuel use by 1.25 kg compared to the control group (p=0.01), demonstrating fuel loading behavior change on the intensive margin. This result complements the finding of Ashraf et al. (2014) that prosocial motivation measured in a charitable dictator game crowds in intrinsic motivation, enhancing the treatment effects of non-financial incentives for public service delivery agents.

iv. Performance and utility

In Table 8, we assess whether job performance is a predictor of job satisfaction.[90] In the theory of captains' behavior in Chapter III, we argue that a captain whose performance exceeds his target will achieve higher utility than a captain who does not achieve his target in accordance with reference dependence theory. To test this conjecture, we aggregated the number of targets met over the course of the experiment for each captain. That is, for those in groups that did not receive a personalized target (i.e. control and information groups), we created 'placebo' targets that mimic those provided to captains in the targets and prosocial incentive groups. For each captain, we summed the number of (hypothetical) targets met during the course of the study and regressed job satisfaction on this job performance proxy variable.

For the groups that received targets in the study (i.e. the targets and prosocial groups), we find that job satisfaction increases by 0.058 points (0.9% effect; p=0.07) on the seven-point scale for each target met during the study; this effect disappears when we restrict the regression to groups who did not receive targets (i.e. the control and information groups). In other words, assuming a linear effect, a captain who met all of his targets (out of a possible 24) would rate his

[89] Importantly, when we examine differences in self-reported giving between 2013 and 2014, we find no evidence that prosocial incentives—either being in the treatment condition, or the extent to which prosocial incentives were achieved—crowd out private charitable giving behavior outside of the study context. In fact, only 27% of captains selected different donation intervals for 2013 and 2014, and 11% claimed to have donated more in 2014.

[90] Captains in Virgin Atlantic express a higher job satisfaction than the average UK citizen, as measured in a random sample who took the BHPS from 1991-2008 (see Dawson, Veliziotis, and Hopkins, 2014). On average, job satisfaction averages 5.36 in the BHPS compared to 5.78 for captains in our survey; average scores according to study group are 5.58 (control), 5.79 (information), 5.82 (targets), and 5.94 (prosocial). While it is not the aim of this paper to assess the influence of preferences on well-being, we find that uncertainty aversion decreases our measure of well-being—job satisfaction—by 0.06 points on the 7-point scale (p=0.038), consistent with the finding of Cardenas and Carpenter (2013) that ambiguity aversion reduces general well-being. Neither the gamble selection nor the altruism proxy influence satisfaction.

overall job satisfaction 1.2 points higher than a captain who did not meet any of his targets (i.e. a 21% improvement on the mean reported job satisfaction). This finding bodes well with the theoretical assumption that performance increases satisfaction, and we interpret this finding as evidence of reference dependence in the context of job performance. Furthermore, since treatment assignment is random, we attribute causality to the relationship between job performance and satisfaction.

To further refine this estimate, we run a similar regression that instead divides the independent variable into three separate variables capturing the number of targets met for each of the three fuel-efficient behaviors. The results of this regression indicate that Efficient Taxi attainment is the driver of the effect of performance on satisfaction in this context, increasing self-reported job satisfaction by 0.12 points for each successful attainment of the monthly target (p=0.03). One potential reason behind this interesting job satisfaction result is that Efficient Taxi is arguably the easiest of the behaviors to achieve and received the most pronounced boost from the treatments; however, we (experimenters, VAA) signal them as equivalently desirable by giving them equal prominence in the feedback forms and the same prosocial incentives across behaviors. Thus, the exogenously induced increase in Efficient Taxi implementation leads to an increase in on-the-job utility for captains who receive targets and achieve them.

Finally, in an era where captains' well-being is especially central to airlines' and travelers' considerations, one might inquire whether the captains themselves are better off. We only take a first step down this important path by considering captains' job satisfaction. Table 9 presents the intent-to-treat estimates for the effects of being in each treatment group on job satisfaction. While we do not find that prosocial incentives improve job satisfaction beyond the effects of targets in isolation, the three components together—information, targets, and prosocial incentives—appear to boost utility in an additively separable manner, culminating in a (weakly) statistically distinguishable improvement for the prosocial incentives group compared to the control group (β=0.37, p=0.105).[91] For context, the difference in self-reported job satisfaction is equivalent to that between an employee with poor health compared to an employee with excellent health (see Clark and Oswald, 1996).

[91] Consistent with the above finding on reference dependence, we also find that captains in the prosocial group who donated more to charity through the study by meeting more of their targets report a higher job satisfaction. Specifically, when we regress job satisfaction on number of targets met in the prosocial group, we find a 0.08-point increase in job satisfaction for each target met (and, consequently, £10 donated to charity on their behalves; p=0.05).

5. Methodological considerations and limitations

Complementary framed (or natural) and artefactual field experiments in a study of this kind can be modified in the following ways. First, one could measure the variables of interest both before and after intervention. Here, for example, it would be useful to have data on captains' preferences prior to the interventions to ensure that these measures were not influenced by the framed field experiment. However, one must consider how the act of taking such measurements may influence subsequent treatment effects, a consideration that precluded our implementing a baseline risk elicitation survey. On a related note, in similar contexts, one could assess employee satisfaction prior to experiment implementation to provide a within-subject measure of changes in utility; here, we (reasonably) assume job satisfaction prior to the study is orthogonal to treatment.

Additionally, participation in our survey was not entirely balanced, though airlines can perhaps overcome this issue by inserting a few short questions in an otherwise mandatory form, for example. Finally, critics may argue that respondents do not have incentive to accurately report their own risk and uncertainty attitudes or donations to charity, especially if they are driven by image concerns. However, regardless of their incentive compatibility, the attitudinal and altruistic measures are clearly predictive of some notable outcomes in this context, and are much less costly to elicit. Moreover, the self-reported risk attitudes have been demonstrated to correlate with incentive-compatible measures in this and aforementioned studies.

6. Discussion

When it comes to the external validity of field experiments, perhaps the most pressing concern is that an average treatment effect in one context is not necessarily generalizable to another. This concern may stem from a number of inconsistencies across populations or contexts, such as differences in population characteristics, political or legal infrastructure, or company standards or culture. Much less do we hear the concern that an average treatment effect may only be capturing the effects on a few, while many others are non-responsive or even respond poorly, as in Costa and Kahn (2013). There has been a significant movement in the direction of understanding heterogeneities of treatment effects, especially with respect to demographic characteristics. Given the low demographic heterogeneity among commercial pilots

in many airlines (e.g., less than six percent of commercial pilots in the entire United States are women[92]) and likely within many organizations, the ability to tailor interventions on observable characteristic may be limited.

Here, we explore a new means by which a principal may seek to tailor information and incentives in a manner that can inform both employee and firm well-being. While economists generally prefer incentive-compatible risk preference elicitation methods for the study of utility functions, we find that self-reported risk attitudes provide more useful information than incentive-compatible measures when predicting performance outcomes under conditions of risk.[93] Using risk attitudes, we find that more risk-averse captains are more inclined to over-fuel and less inclined to accurately implement the targeted Fuel Load measure. Attitudes toward uncertainty are less predictive of general fuel loading efficiency, and altruism appears to be orthogonal.

However, altruism—as proxied by annual donations prior to the study—is associated with a stronger treatment effect of prosocial incentives. Captains with a higher revealed preference for charitable donations in their private lives are also more inclined to implement the prosocially incentivized behavior. Interestingly, more altruistic captains respond more strongly to non-financial incentives—i.e. exogenous target provision—on the intensive margin, reducing fuel use relative to the optimal, even if doing so does not trigger a success on the binary outcome encouraged in the study. A plausible interpretation of this phenomenon is that incentive provision places salience on the incentivized behavior itself, whereas the softer targets intervention places more emphasis on the importance of the outcome of interest (i.e. fuel use).

Finally, with respect to captain welfare, we find that provision of targets and prosocial incentives can increase employee satisfaction if captains are able to meet their targets, indicating some degree of reference dependence along this dimension. Target provision may, therefore, provide a highly cost-effective means for an airline to both improve job performance (i.e. fuel efficiency) and improve employee satisfaction, perhaps especially if captains are provided with additional support or training that allows them to meet their targets consistently. Adding small prosocial incentives layered on top of personalized targets may not lead to increased performance

[92] "US Civil Airmen Statistics." 2015. <https://www.faa.gov/data_research/aviation_data_statistics/civil_airmen_statistics/>.

[93] As briefly mentioned in the introduction, a vast literature has argued and provided evidence for selection of workers into contracts based on risk preferences; therefore, it is possible that homogeneity of latent risk preferences among the sample under scrutiny here may prevent high-powered estimation of the effects of risk on performance outcomes. Nevertheless, if such selection exists and therefore heterogeneity of risk preferences is minimal, the effect of such latent preference heterogeneity on performance is irrelevant.

across the board (as shown in Chapter III), though it may do so for those captains exhibiting high levels of altruism exhibited through their private charitable giving. Additionally, merely receiving the prosocial incentives feedback improved captains' average job satisfaction relative to the control group at statistically discernible levels.

In sum, there may be much we can learn from heterogeneities not only in demographic characteristics, but also in self-reported or revealed preferences. Future field experimental research may consider complementing interventions with surveys rooted in preference and attitude elicitation so that we may further understand the role of preference and personality traits in outcome variables of interest. In this way, we can seek to provide individuals and businesses with feedback and incentives that not only improve outcomes that decision makers in the policy and business worlds may care about, but also improve the well-being of the targeted unit of analysis, ensuring that both efficiency and welfare objectives are achieved.

REFERENCES

Ackerberg, Daniel A., and Maristella Botticini. 2002. "Endogenous matching and the empirical determinants of contract form." *Journal of Political Economy* 110 (3): 564-591.

Allen, Douglas W., and Dean Lueck. 1992. "Contract choice in modern agriculture: Cash rent versus cropshare." *The Journal of Law and Economics* 35 (2): 397-426.

Allen, Douglas W., and Dean Lueck. 1995. "Risk preferences and the economics of contracts." *American Economic Review* 85 (2): 447-451.

Allen, Douglas W., and Dean Lueck. 1999. "The role of risk in contract choice." *Journal of Law, Economics, and Organization* 15 (3): 704-736.

Andersen, Steffen, Glenn W. Harrison, Morten Igel Lau, and E. Elisabet Rutström. 2006. "Elicitation using multiple price list formats." *Experimental Economics* 9 (4): 383-405.

Andreoni, James. 1989. "Giving with impure altruism: Applications to charity and Ricardian equivalence." *The Journal of Political Economy* 97 (6): 1447-1458.

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. 2014. "No margin, no mission? A field experiment on incentives for public service delivery." *Journal of Public Economics* 120: 1-17.

Ashton, Michael C. 1998. "Personality and job performance: The importance of narrow traits." *Journal of Organizational Behavior* 19: 289-303.

Barr, Abigail, and Garance Genicot. 2008. "Risk sharing, commitment, and information: An experimental analysis." *Journal of the European Economic Association* 6 (6): 1151-1185.

Barsky, Robert B., F. Thomas Juster, Miles S. Kimball, and Matthew D. Shapiro. 1997. "Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study." *Quarterly Journal of Economics* 112 (2): 537-579.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak. 1964. "Measuring utility by single-response sequential method." *Behavioral Science* 9 (3): 226-232.

Bellante, Don, and Albert N. Link. 1981. "Are public sector workers more risk averse than private sector workers?" *Industrial and Labor Relations* 34 (3): 408-412.

Bellemare, Charles, and Bruce Shearer. 2010. "Sorting, incentives, and risk preferences: Evidence from a field experiment." *Economics Letters* 108 (3): 345-348.

Bellemare, Charles, and Bruce Shearer. 2013. "Multidimensional heterogeneity and the economic importance of risk and matching: Evidence from contractual data and field experiments." *RAND Journal of Economics* 44 (2): 361-389.

Biddle, Jeff E., and Gary A. Zarkin. 1988. "Worker preference and market compensation." *The Review of Economics and Statistics* 70 (4): 660-667.

Binswanger, Hans P. 1980. "Attitudes toward risk: Experimental measurement in rural India." *American Journal of Agricultural Economics* 62 (2): 395-407.

Bonin, Holger, Thomas Dohmen, Armin Falk, David Huffman, and Uwe Sande. 2007. "Cross-sectional earnings risk and occupational sorting: The role of risk attitudes." *Labour Economics* 14 (6): 926-937.

Bougherara, Douadia, Xavier Gassmann, and Laurent Piet. 2011. "Eliciting risk preferences: A field experiment on a sample of French farmers." *2011 International Congress, Zurich, Switzerland*. No. 114266. European Association of Agricultural Economists.

Brock, J. Michelle, Andreas Lange, and Kenneth L. Leonard. 2016. "Generosity and prosocial behavior in healthcare provision: Evidence from the laboratory and field." *Journal of Human Resources* 51 (1): 133-162.

Burks, Stephen V., Daniele Nosenzo, Jon Anderson, Matthew Bombyk, Derek Ganzhorn, Lorenz Götte, and Aldo Rustichini. 2016. "Lab measures of other-regarding preferences can predict some related on-the-job behavior: Evidence from a large scale field experiment." IZA Working Paper 9767.

Buurman, Margaretha, Josse Delfgaauw, Robert Dur, and Seth Van den Bossche. 2012. "Public sector employees: Risk averse and altruistic?" *Journal of Economic Behavior and Organization* 83 (3): 279-291.

Callen, Michael, Saad Gulzar, Ali Hasanain, Yasir Khan, and Arman Rezaee. 2015. "Personalities and public sector performance: Evidence from a health experiment in Pakistan." *National Bureau of Economic Research* Working Paper No. w21180.

Cardenas, Juan Camilo, and Jeffrey Carpenter. 2013. "Risk attitudes and economic well-being in Latin America." *Journal of Development Economics* 103: 52-61.

Charness, Gary, and Angelino Viceisza. 2016. "Three risk-elicitation methods in the field: Evidence from rural Senegal." *Review of Behavioral Economics* 3 (2): 145-171.

Charness, Gary, Uri Gneezy, and Alex Imas. 2013. "Experimental methods: Eliciting risk preferences." *Journal of Economic Behavior and Organization* 87: 43-51.

Clark, Andrew E., and Andrew J. Oswald. 1996. "Satisfaction and comparison income." *Journal of Public Economics* 61 (3): 359-381.

Cornelissen, Thomas, John S. Heywood, and Uwe Jirjahn. 2011. "Performance pay, risk attitudes and job satisfaction." *Labour Economics* 18 (2): 229-239.

Costa, Dora L. and Matthew E. Kahn. 2013. "Energy conservation 'nudges' and environmentalist ideology: Evidence from a randomized residential electricity field experiment." *Journal of the European Economic Association* 11 (3): 680-702.

Coval, Joshua D., and Tyler Shumway. 2005. "Do behavioral biases affect prices?" *The Journal of Finance* 60 (1): 1-34.

Dasgupta, Utteeyo, Subha Mani, Smriti Sharma, and Saurabh Singhal. 2016. "Eliciting risk preferences: Firefighting in the field." IZA Discussion Paper No. 9765.

Dave, Chetan, Catherine C. Eckel, Cathleen A. Johnson, and Christian Rojas. 2010. "Eliciting risk preferences: When is simple better?" *Journal of Risk and Uncertainty* 41 (3): 219-243.

Dawson, Chris, Michail Veliziotis, and Benjamin Hopkins. 2014. "Temporary employment, job satisfaction and subjective well-being." *Economic and Industrial Democracy* (2014), doi: 10.1177/0143831X14559781.

Di Mauro, Carmela, and Rosy Musumeci. 2011. "Linking risk aversion and type of employment." *The Journal of Socio-Economics* 40 (5): 490-495.

Dizon-Ross, Rebecca, Pascaline Dupas, and Jonathan Robinson. 2015. "Governance and the effectiveness of public health subsidies." *National Bureau of Economic Research* Working Paper No. w21324.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual risk attitudes: Measurement, determinants, and behavioral consequences." *Journal of the European Economic Association* 9 (3): 522-550.

Eckel, Catherine C., and Philip J. Grossman. 2008. "Forecasting risk attitudes: An experimental study using actual and forecast gamble choices." *Journal of Economic Behavior and Organization* 68 (1): 1-17.

Eil, David, and Jaimie W. Lien. 2014. "Staying ahead and getting even: Risk attitudes of experienced poker players." *Games and Economic Behavior* 87: 50-69.

Elston, Julie A., Glenn Harrison, and Elisabet Rutström. 2006. "Experimental economics, entrepreneurs and the entry decision." *University of Central Florida working paper*.

Falco, Paolo. 2014. "Does risk matter for occupational choices? Experimental evidence from an African labour market." *Labour Economics* 28: 96-109.

Fouarge, Didier, Ben Kriechel, and Thomas Dohmen. 2014. "Occupational sorting of school graduates: The role of economic preferences." *Journal of Economic Behavior and Organization* 106: 335-351.

Groom, Ben, Phoebe Koundouri, Céline Nauges, and Alban Thomas. 2008. "The story of the moment: Risk averse Cypriot farmers respond to drought management." *Applied Economics* 40 (3): 315-326.

Grund, Christian, and Dirk Sliwka. 2010. "Evidence on performance pay and risk aversion." *Economics Letters* 106 (1): 8-11.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42 (4): 1009-1055.

Harrison, Glenn W., and E. Elisabet Rustrom. 2008. "Risk aversion in the laboratory." *Research in Experimental Economics* 12: 41-196.

Herzog, Jr., Henry W., and Alan M. Schlottmann. 1990. "Valuing risk in the workplace: Market price, willingness to pay, and the optimal provision of safety." *The Review of Economics and Statistics* 72 (3): 463-470.

Hey, John D., and Chris Orme. 1994. "Investigating generalizations of expected utility theory using experimental data." *Econometrica* 62 (6): 1291-1326.

Holmström, Bengt. 1979. "Moral hazard and observability." *The Bell Journal of Economics* 10 (1): 74-91.

Holmström, Bengt, and Paul Milgrom. 1991. "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." *Journal of Law, Economics, and Organization* 7: 24-52.

Holt, Charles A., and Susan K. Laury. 2002. "Risk aversion and incentive effects." *American Economic Review* 92 (5): 1644-1655.

Ihli, Hanna Julia, Brian Chiputwa, and Oliver Musshoff. 2016. "Do changing probabilities or payoffs in lottery-choice experiments affect risk preference outcomes? Evidence from rural Uganda." *Journal of Agricultural and Resource Economics* 41 (2): 324-345.

Jacobsen, Sarah and Reagan Petrie. 2009. "Learning from mistakes: What do inconsistent choices over risk tell us?" *Journal of Risk and Uncertainty* 38: 143-158.

Jamison, Julian, Dean Karlan, and Jonathan Zinman. 2012. "Measuring risk and time preferences and their connections with behavior." Unpublished manuscript.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect theory: An analysis of decision under risk." *Econometrica* 47 (2): 263-292.

King, Allan G. 1974. "Occupational choice, risk aversion, and wealth." *Industrial and Labor Relations Review* 27 (4): 586-596.

Kumbhakar, Subal C., and Ragnar Tveterås. 2003. "Risk preferences, production risk, and firm heterogeneity." *The Scandinavian Journal of Economics* 105 (2): 275-293.

Lafontaine, Francine, and Sugato Bhattacharyya. 1995. "The role of risk in franchising." *Journal of Corporate Finance* 2 (1): 39-74.

Masclet, David, Nathalie Colombier, Laurent Denant-Boemont, and Youenn Lohéac. 2009. "Group and individual risk preferences: A lottery-choice experiment with self-employed and salaried workers." *Journal of Economic Behavior and Organization* 70: 470-484.

Picazo-Tadeo, Andrés, and Alan Wall. 2011. "Production risk, risk aversion and the determination of risk attitudes among Spanish rice producers." *Agricultural Economics* 42 (4): 451-464.

Prendergast, Canice. 2002. "The tenuous trade-off between risk and incentives." *The Journal of Political Economy* 110 (5): 1071-1102.

Reynaud, Arnaud, and Stéphane Couture. 2012. "Stability of risk preference measures: Results from a field experiment on French farmers." *Theory and Decision* 73 (2): 203-221.

Rustichini, Aldo, Colin G. DeYoung, Jon Anderson, and Stephen V. Burks. 2016. "Toward the integration of personality theory and decision theory in explaining economic behavior: An experimental investigation." *Journal of Behavioral and Experimental Economics* 64: 122-137.

Ryerson, Megan S., Mark Hansen, and Michael Seelhorst. 2015. "Landing on empty: Estimating the benefits from reducing fuel uplift in US Civil Aviation." *Environmental Research Letters* 10: 1-11.

Scholz, John T., and Wayne B. Gray. 1990. "OSHA enforcement and workplace injuries: A behavioral approach to risk assessment." *Journal of Risk and Uncertainty* 3: 283-305.

Shaw, Kathryn L. 1996. "An empirical analysis of risk aversion and income growth." *Journal of Labor Economics* 14 (4): 626-653.

Smith, V. Kerry, Subhrendu K. Pattanayak, and George L. Van Houtven. 2003. "VSL reconsidered: What do labor supply estimates reveal about risk preferences?" *Economics Letters* 80 (2): 147-153.

Stiglitz, Joseph E. 1974. "Incentives and risk sharing in sharecropping." *The Review of Economic Studies* 41 (2): 219-255.

Takahashi, Kazushi. 2013. "The roles of risk and ambiguity in the adoption of the system of rice intensification (SRI): Evidence from Indonesia." *Food Security* 5 (4): 513-524.

Tanaka, Tomomi, Colin F. Camerer, and Quang Nguyen. 2010. "Risk and time preferences: Linking experimental household survey data from vietnam." *American Economic Review* 100 (1): 557-571.

Tett, Robert P., Douglas N. Jackson, and Mitchell Rothstein. 1991. "Personality measures as predictors of job performance: A meta-analytic review." *Personnel Psychology* 44 (4): 703-742.

Viscusi, W. Kip, and Joni Hersch. 2001. "Cigarette smokers as job risk takers." *The Review of Economics and Statistics* 83 (2): 269-280.

Webster, Donna M., and Arie W. Kruglanski. 1994. "Individual differences in need for cognitive closure." *Journal of Personality and Social Psychology* 67 (6): 1049-1062.

TABLES

TABLE 1
RISK PREFERENCE ELICITATION: GAMBLE CHARACTERISTICS

	Probabilities	Event A	Event B	Expected Payoff	Risk	CRRA Range
Gamble 1	50/50	£25	£25	£25	0	$r>2.49$
Gamble 2	50/50	£45	£15	£30	15	$0.84<r<2.49$
Gamble 3	50/50	£65	£5	£35	30	$0.50<r<0.84$
Gamble 4	50/50	£85	-£5	£40	45	$0.33<r<0.50$
Gamble 5	50/50	£105	-£15	£45	60	$r<0.33$

Notes: The CRRA range is calculated including the survey endowment of £25. The ranges represent the *r* levels for which a captain with CRRA utility would select the given gamble.

TABLE 2
BALANCE CHECK I: CAPTAIN-LEVEL OBSERVABLES

	U: Unopened (N=133)	S: Started (N=202)	Test of equality: U vs. S	I: Incomplete (N=142)	C: Complete (N=193)	Test of Equality: I vs. C	Test
Age (years)	52.03	51.86	p=0.778	52.06	51.83	p=0.695	t-test
Seniority	157.55	178.74	p=0.052	157.41	179.83	p=0.038	t-test
Flights	131.80	130.76	p=0.811	131.43	130.98	p=0.917	t-test
Fuel Load	0.401	0.458	p=0.004	0.405	0.457	p=0.009	t-test
Trainer	0.150	0.208	p=0.185	0.169	0.197	p=0.516	χ^2
Trusted Pilot	0.015	0.045	p=0.138	0.021	0.041	p=0.302	χ^2

Notes: A total of 202 subjects started the survey, and 193 completed it. Those who started answered the question pertaining to job satisfaction but did not make enough progress in the survey to answer the risk, ambiguity, or social preference questions. The above represent p-values for chi-square tests for differences of sample proportions across various demographic and occupational variables. "Flights" captures the number of flights flown from January 2013 through March of 2015 (the period captured in our dataset), though the test results are unchanged when we solely look at flights in the pre-experimental and experimental periods in isolation.

TABLE 3
BALANCE CHECK II: STUDY GROUP ASSIGNMENT

	C: Control	I: Info	Test: C vs. I	T: Targets	Test: C vs. T	Test: I vs. T	P: Prosocial	Test: C vs. P	Test: I vs. P	Test: T vs. P
Started	41.2%	37.6%	0.638	44.4%	0.671	0.373	35.7%	0.466	0.794	0.252
Complete	44.7%	40.0%	0.535	45.7%	0.900	0.460	39.2%	0.475	0.924	0.406

Notes: A total of 202 subjects started the survey, and 193 completed it. Those who started answered the question pertaining to job satisfaction but did not make enough progress in the survey to answer the risk, ambiguity, or social preference questions. The above represent p-values for chi-square tests for differences of sample proportions across the experimental study groups. The above test is based on all project flights from January 2013 through March of 2015, though the test results are unchanged when we solely look at flights in the pre-experimental and experimental periods in isolation.

TABLE 4
RISK PREFERENCES AND RISK ATTITUDES

Gamble Selection	Respondents	Risk Attitude: Mean (SD)
1	44 (22.8%)	3.23 (1.88)
2	29 (15.0%)	4.86 (2.49)
3	45 (23.3%)	5.20 (2.15)
4	25 (13.0%)	5.16 (2.17)
5	50 (25.9%)	4.58 (2.19)
Total	193 (100%)	4.53 (2.26)

TABLE 5
SURVEY VARIABLE NAMES AND DESCRIPTIONS

Variable Name	**Description**	**Source**	**Values**
Risk_pref	Incentive-compatible risk preference	Binswanger (1980); Barr and Genicot (2008); Eckel and Grossman (2008)	Integer from 1 (most risk averse) to 5 (most risk loving)
Risk_attitude	Self-reported risk attitude	German Socio-Economic Panel	Integer from 0 (not at all willing to take risks) to 10 (very willing to take risks)
Unc_averse	Self-reported uncertainty aversion	Need for Closure Scale (Webster and Kruglanski, 1994)	Integer from 0 (most averse) to 10 (least averse)
Donations	Self-reported donations to charity in 2013	-	Midpoint of interval selected (of which there are ten), x10 for ease of regression interpretation
Job Satisfaction	Self-reported job satisfaction	British Household Panel Survey	Integer from 1 (lowest) to 7 (highest)

TABLE 6

RISK, UNCERTAINTY, AND FUEL EFFICIENCY

	(1) Fuel Load	(2) Fuel Diff	(3) Fuel Load	(4) Fuel Diff	(5) Fuel Load	(6) Fuel Diff
Risk_pref	0.003	-22.044				
	(0.007)	(14.808)				
Risk_attitude			0.009**	-16.803*		
			(0.004)	(8.826)		
Unc_averse					-0.003	14.029*
					(0.004)	(8.136)
Constant	0.149	1,542.758*	0.062	1,597.365*	0.176	1,360.372
	(0.349)	(861.115)	(0.339)	(853.709)	(0.351)	(876.394)
Observations	11,770	11,770	11,770	11,770	11,770	11,770
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The dependent variables in these regressions indicate the proportion of flights for which Fuel Load was successfully performed ('Fuel Load') and the continuous difference between actual fuel uptake and the 'correct' fuel uptake determined by the zero fuel weight calculation ('Fuel Difference'). Robust standard errors (clustered at the captain level) are reported below estimates in parentheses. Total flight observations are provided. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, whether the captain has completed training, and observed captain characteristics (age, seniority, and dummies for whether the pilot is a trainer or 'trusted pilot'). ***p<0.01 **p<0.05 *p<0.10

TABLE 7

SOCIAL PREFERENCES AND FUEL EFFICIENCY

	(1)	(2)	(3)	(4)
	Fuel Load	Fuel Diff	Fuel Load	Fuel Diff
Donations	0.001	-3.206	0.001	-1.824
	(0.001)	(2.092)	(0.001)	(2.537)
Expt			0.049*	-165.053***
			(0.028)	(60.867)
Information			-0.022	-16.918
			(0.025)	(54.225)
Targets			-0.010	-170.557***
			(0.025)	(55.135)
Prosocial			-0.090***	143.366***
			(0.023)	(53.656)
Expt*Information			0.001	57.452
			(0.039)	(81.959)
Expt*Targets			0.025	61.783
			(0.040)	(78.574)
Expt*Prosocial			-0.053	86.836
			(0.036)	(77.805)
Expt*Don13			-0.000	0.314
			(0.000)	(0.347)
Information*Donations			0.000	-0.143
			(0.000)	(0.308)
Targets*Donations			-0.000	0.854**
			(0.000)	(0.358)
Prosocial*Donations			0.000**	-0.694**
			(0.000)	(0.326)
Expt*Info*Donations			0.000	-0.278
			(0.000)	(0.449)
Expt*Targets*Donations			0.000	-1.247***
			(0.000)	(0.485)
Expt*Prosocial*Donations			0.001***	-0.608
			(0.000)	(0.470)
Constant	0.293	384.741	0.302	546.635
	(0.286)	(722.159)	(0.255)	(652.052)
Observations	11,434	11,434	18,776	18,776
Controls	Yes	Yes	Yes	Yes

Notes: The dependent variables in these regressions indicate the proportion of flights for which Fuel Load was successfully performed ('Fuel Load') and the continuous difference between actual fuel uptake and the 'correct' fuel uptake determined by the zero fuel weight calculation ('Fuel Diff'). Columns (1) and (2) show the results of an OLS regression of the dependent variables on altruistic preferences in the pre-experimental period alone (i.e. before any experimental interventions were introduced). Columns (3) and (4) show the results of a difference-in-difference regression specification comparing pre-experiment behavior (January 2013-January 2014) to behavior during the experiment (February 2014-September 2014; "Expt"); as such, the coefficients indicate the increase in the proportion of flights beyond the control group for which the behavior of interest was successfully performed. Newey-West standard errors (lag = 1) are reported below estimates in parentheses. Total flight observations are provided. The covariate specific to this regression is captains' pro-social behavior proxied by self-reported donations in 2013 (Don13 captures the midpoints of ten donation amount intervals). We divide these midpoints by ten so that Don13 captures the effects of increases of £10 in personal donations. Controls include weather on departure and arrival, number of engines on the aircraft, aircraft type, ports of departure and arrival, aircraft maintenance, captains' contracted hours, and whether the captain has completed training. ***p<0.01 **p<0.05 *p<0.10

TABLE 8

JOB SATISFACTION AND JOB PERFORMANCE

	Groups: Control and Info		Groups: Targets and Charity	
Targets Met:	Job Satisfaction		Job Satisfaction	
Fuel Load	0.093	-	0.065	-
	(0.062)		(0.060)	
Efficient Flight	-0.074	-	-0.017	-
	(0.056)		(0.054)	
Efficient Taxi	0.025	-	0.120**	-
	(0.043)		(0.054)	
Overall	-	0.006	-	0.058*
		(0.028)		(0.031)
Constant	5.691***	5.632***	5.399***	5.33***
	(0.291)	(0.300)	(0.358)	(0.358)
N	*N*=103 subjects		*N*=99 subjects	
Obs	103		99	
Controls	None		None	

Notes: The dependent variable in these regressions is a 7-point scale of job satisfaction, where self-reported job satisfaction increases in the scale. Robust standard errors are reported below estimates in parentheses. The independent variables indicate the number of targets met per behavior as well as overall during the course of the study. ***p<0.01 **p<0.05 *p<0.10

TABLE 9

JOB SATISFACTION AND TREATMENT ASSIGNMENT

	Job Satisfaction
TG1: Information	0.212
	(0.224)
TG2: Targets	0.242
	(0.249)
TG3: Prosocial	0.365
	(0.223)
Constant	5.58***
	(0.174)
Observations	202
Controls	None

Notes: The dependent variable in this regression is a 7-point scale of job satisfaction, where self-reported job satisfaction increases in the scale. Robust standard errors are reported below estimates in parentheses.

APPENDIX: SURVEY MATERIALS

Risk preference elicitation question

Below you will be asked to select from among five different gambles the one gamble you would like to play. The five different gambles are listed below.

Each gamble has two possible outcomes, both of which are associated with a 50% probability of payout (i.e. a fair coin toss).

Please note that if you should select either Gamble 4 or Gamble 5 and you incur a loss (i.e. -£5 or -£15), your losses will be deducted from your £25 compensation for completing the survey (e.g., £25 - £5 = £20).

Please select your preferred 50/50 gamble.

- ○ **Gamble 1:** £25 vs. £25
- ○ **Gamble 2:** £45 vs. £15
- ○ **Gamble 3:** £65 vs. £5
- ○ **Gamble 4:** £85 vs. -£5
- ○ **Gamble 5:** £105 vs. -£15

CHAPTER V

A BARGAINING EXPERIMENT ON HETEROGENEITY AND SIDE DEALS IN CLIMATE NEGOTIATIONS

By Greer Gosnell and Alessandro Tavoni

Abstract: The recent global climate change agreement in Paris leaves a wide gap between pledged and requisite emissions reductions in keeping with the commonly accepted 2°C target. A recent strand of theoretical and experimental evidence establishes pessimistic predictions concerning the ability of comprehensive global environmental agreements to improve upon the business-as-usual trajectory. We introduce an economic experiment focusing on the dynamics of the negotiation process by observing subjects' behavior in a Nash bargaining game. Throughout repeated rounds, heterogeneous players bargain over the allocation of a fixed amount of profit-generating emissions with significant losses attached to prolonged failure to reach agreement. We find that the existence of side agreements that constrain individual demands among a subset of like countries does not ensure success; however, such side agreements reduce the demands of high-emission parties. Our results highlight the importance of strong signals amongst high emitters in reaching agreement to shoulder a collective emissions reduction target.

Acknowledgments: Financial support by Enel Foundation (Grant 1-RGI-U885) is kindly acknowledged. This work is part of the program of the Centre for Climate Change Economics and Policy, which is funded by the UK Economic and Social Research Council. Special thanks to Jeroen Nieboer and the Behavioural Research Lab for assisting with the software and the laboratory facilities.

1. Introduction

Recent developments in climate policy have reaffirmed the perceived importance of minilateral agreements among a small number of countries prior to engaging in large *fora* such as the annual Conferences of the Parties (COPs). A growing literature in political science points to the merits and drawbacks of entering into negotiations among small-n clubs (1-3). At the two ends of the spectrum, one finds bilateral negotiations and almost universal groupings like the United Nations Framework Convention on Climate Change COPs. Most experts agree that bottom-up and top-down approaches are not mutually exclusive (4, 5). Indeed, it appears that some countries have resorted to bilateral deals as a stimulus for action by less motivated countries in global negotiations, a common reading of the U.S.-China joint announcement to reduce emissions that took place ahead of the 21st COP in Paris. The pledges in the announcement were cemented in the countries' Intended Nationally Determined Contributions (INDCs) (6).

Would countries commit to emissions cuts if assured of others' intentions to invest in climate change mitigation? This question is of course an empirical one, and its answer hinges on the success of ongoing international climate negotiations and the ensuing burden-sharing settlement. However, it will take years before the implications of such agreements can be (imprecisely) quantified in terms of emissions reductions. In the meantime, one may approach the issue with other tools, such as theoretical modeling and laboratory experimentation. Inspired by a bargaining model that aims to capture some of the stylized tradeoffs inherent in climate change negotiations (7), we introduce a novel economic experiment that focuses on the role of side deals reached by a subset of negotiators in shaping subsequent global negotiations.

Smead and coauthors (7) use an agent-based model with learning dynamics to examine past failures and future prospects for an international climate agreement. In the model, agents play an N-player Nash bargaining game (8-10), where each player's strategy set is the interval [0,1] representing the range of possible reductions: 1 constituting business-as-usual (BAU) and 0 constituting a complete reduction to zero emissions. In addition to imposing learning dynamics, they modify the Nash bargaining game by introducing an exogenous global emissions target T in the interval (0,1). Players maintain the full amount demanded from the global "emissions pie"—where a higher share translates to a higher payoff—only if the sum of all individual demands does not exceed the targeted proportion of BAU emissions (and receive a small fraction δ of their demands otherwise). The authors vary a number of parameters in the model and find that player

heterogeneity increases the likelihood of success, and that prior minilateral agreements can facilitate collective agreement (especially those made among a large number of small players as opposed to a small number of large players, *ceteris paribus*).

We explore this issue of negotiating on costly emissions reductions in the laboratory. The experimental literature on the avoidance of dangerous climate change has thus far focused on the provision of threshold public goods (11-16). The underlying idea is that, in order to stay within a safe operating space and avoid probabilistic losses, players must invest sufficient resources into a public account (17-22). One can view this public good as a minimum collective expenditure in climate change mitigation that ensures staying below an agreed temperature change, such as the often-mentioned 2°C target.

Since climate negotiations entail agreement on emissions reductions with a view toward remaining within a given threshold, we instead frame the costly mitigation problem as a modified Nash bargaining game. This approach has thus far been neglected in the experimental literature on climate change cooperation. In the game, payoffs accrue only if the groups' demands fall within a given threshold of available emissions. Negotiators must divide the burden of reducing the size of the emissions pie by agreeing on sufficiently ambitious reductions relative to BAU, which in the game is represented by players' initial endowments. The underlying assumption is that emissions map one-to-one with wealth. While this assumption is undoubtedly a strong simplification of complex dynamics, it allows us to isolate important features of climate change negotiations, such as the tension between a country's incentive to keep the largest possible fraction of its emissions and the need to make concessions if the collective target is to be met. That is, future emissions reductions generally bear significant opportunity costs in terms of burdens associated with compliance. Since historical responsibilities are not explicitly modeled, the correlation simply aims to capture the pervasive notion of economic sacrifice on the part of countries that commit to future emissions reductions.

In addition to the experimental methodology employed, we depart from (7) in two noteworthy ways. First, in our design, the loss incurred by a group that fails to reach agreement is independent of individual demands. This feature is consistent with the standard bargaining game formulation, which prescribes that out-of-equilibrium payoffs are constant. More importantly, to capture the realistic feature that delay in reaching agreement over ambitious emissions reductions will result in the need to agree on even more ambitious targets in the future, we designed the

game to comprise multiple rounds with increasingly stringent targets (see Figure 1). Hence, while selfish motives still push in the direction of high demands in the hope that others will lead the effort, there is a critical urgency for the negotiating group to meet its target.

Strategic implications of costly haggling, i.e. costs associated with delay in reaching agreement, have been studied extensively. The alternating-offers model entails the partition of a cake between a proposer and a second mover (23). If the latter rejects the offer, she becomes the proposer and the process is repeated. This alternation of roles continues until an agreement is struck, at which point the cake is divided accordingly. The game-theoretic solution predicts instantaneous agreement on the division of the cake, with the proposer securing a weakly larger share, depending on the discount factor. The game analyzed here differs along the following dimensions: number of players (we focus on multilateral bargaining); timing of the proposals (negotiators move simultaneously); horizon (players have a finite number of rounds to reach an agreement); and disagreement costs. In the alternating-offers model, costs of inaction arise with the first rejection, and can be thought of as (partial) spoiling of the cake: in the limit, if both players perpetually disagree, their payoffs vanish. Here, the losses are not smooth over time, as is evident from Figure 1. Furthermore, players do not bargain on the status quo, as the climate change problem requires agreement on shrinking the cake from the outset.

Our bargaining game also relates to the ultimatum game, the simplest form of the alternating-offers model where only the final two stages are considered. Hence the ultimatum: the responder's choice is again confined to acceptance or rejection of the offer, with rejection implying a 0 payoff for both players. Under complete information, the subgame perfect Nash equilibrium involves a rational self-interested proposer offering nothing (or an arbitrarily small share) and the responder accepting. However, nontrivial offers have been consistently found in experimental settings due to the proposers' concerns for fairness and fear of rejection of offers below an acceptable threshold (24, 25). In common with the above, our game centers upon issues of burden sharing that are likely to trigger fairness considerations. However, the multilateral and simultaneous nature of the repeated negotiations we simulate in the lab—coupled with the introduction of a target requiring coordination—introduces additional considerations, such as group-level efficiency and reputation. We further explain the implications of the lab design features and discuss its equilibria as well as its relation to the experimental literature in parts (a) and (b) of the Supplementary Information (SI).

2. Methods

In the experiment, groups of six "Countries" negotiate over a maximum of eight rounds on increasingly ambitious collective emissions reduction targets. In each round of negotiation, Countries individually demand to keep a proportion of their endowed (BAU) emissions with the shared group goal of shrinking the global pie in accordance with the exogenous global reduction target.

Each treatment consists of up to eight rounds of a Nash bargaining game framed as a climate change negotiation, where the negotiation terminates if the group meets the prescribed Global Target T in a given round. The Global Target becomes more difficult to attain as the game progresses, beginning at T=60% of global wealth and reducing by 10% every two rounds (i.e. T=50% in Rounds 3-4, T=40% in Rounds 5-6, and T=30% in Rounds 7-8). If the group does not meet the target by the end of Round 8, negotiation terminates and group members each receive δ=10% of their initial endowment (regardless of their demands in the final round) as an unavoidable consequence of "dangerous" climate change.

In every round, group members—each acting as a delegate representing one Country in the negotiation—engage in what we term the Global Negotiation stage. In this stage, each delegate demands to keep a proportion of her Country's endowed emissions, which is perfectly correlated with its wealth in the game. If the group's aggregate demand does not exceed the corresponding Global Target in a given round, the target is met and each subject in the group receives the proportion she demanded in that round. If the target is not met, there is no payout for the round and negotiations continue to the next round.

We implement five variants of the bargaining experiment: Symmetric (SYM), Asymmetric (ASYM), Poor Side Deals (PSD), Rich Side Deals (RSD), and All Side Deals (ASD).[94] All groups' aggregate monetary endowments are £100 (approximately US$156). In treatment SYM, all Countries begin with a symmetric endowment of £16.67. All other treatments are characterized by asymmetry in the distribution of endowments (and corresponding impact on global emissions). In these treatments, four Poor Country delegates each receive an endowment of £10 and two Rich Country delegates each receive an endowment of £30 (see Table 1).

[94] A total of 336 subjects participated in 20 experimental sessions. Eleven groups participated in SYM, 14 in ASYM , 10 in PSD, 10 in RSD, and 11 in ASD.

All treatment conditions consist of eight rounds of negotiation. Treatments without Side Deals—SYM and ASYM—feature only single-stage rounds, as depicted in Figure 1. In each of these rounds, delegates independently and simultaneously decide on individual (i.e. Country-level) demands. The software computes the aggregated 'global' demand of the group and displays both global and individual demands in a subsequent screen in absolute and percentage terms.

In treatments containing Side Deals (PSD, RSD, and ASD), either one or two subsets of delegates—belonging to the same wealth/emissions category, i.e. Poor, Rich, or Poor and Rich, respectively—may collectively place binding constraints on own individual demands in the two upcoming Global Negotiation stages. Accordingly, these Side Deals take place prior to the Global Negotiation stages of Rounds 1, 3, 5, and 7. The outcome of a Side Deal—the *Agreed Maximum Demand*—applies only to Countries who took part in the Side Deal, though it is revealed to all Countries within the group prior to the subsequent Global Negotiation stages. The Agreed Maximum Demand is the mean of the *Maximum Demands*, i.e. the answers of the Side Deal participants to the following question (in the PSD treatment): "What is the maximum percentage of emissions/wealth that you think is appropriate for each Poor Country to demand in each of the two upcoming global negotiations?"

To be clear, we provide the following hypothetical example of Side Deal implementation in the PSD treatment. Prior to the Global Negotiation stage of Round 1, all four Poor Countries will determine an Agreed Maximum Demand, which is a binding constraint on the Poor Countries' individual demands in the Global Negotiation stages of Rounds 1 and 2. In this Side Deal stage, if two Poor Countries choose a Maximum Demand of 80 and two choose a Maximum Demand of 60, the resulting Agreed Maximum Demand is $(2\times60 + 2\times80) \div 4 = 70$. Poor Countries may then only individually demand to keep up to 70% of their own initial endowment in the Global Negotiation stages of Rounds 1 and 2. If the group collectively fails to reach the Global Target of 60% of global wealth/emissions by the end of Round 2, Poor Countries will again enter a Side Deal stage and similarly determine a new Agreed Maximum Demand that pertains to the Global Negotiation stages of Rounds 3 and 4—when the Global Target is reduced to 50%—and so on.[95]

[95] See SI for further details, including Screenshots 4-8 for visual representations of the above material as displayed in the experimental instructions.

3. Results

i. Global success

Asymmetry and side deals. Table 2 provides a descriptive overview of group performance dynamics across treatments. First, we see that *all* symmetric groups had reached agreement by the end of the fourth round of negotiations. When comparing success rates within the first four rounds, the SYM groups outperform the ASYM (proportion test, p=0.101, z=1.64), RSD (proportion test, p=0.049, z=1.96), and ASD (proportion test, p=0.062, z=1.86) groups. This finding is in contrast to the results in (7), where the authors find that asymmetry of endowments increases the likelihood of agreement. A second, more relevant finding is the limited impact of Side Deals on negotiation outcomes. When comparing ASYM groups to all groups containing Side Deals (both pairwise and combined), we do not find conclusive evidence that treatments containing Side Deals improve upon global negotiations that occur among asymmetric actors in the absence of Side Deals, in terms of both agreement velocity and (individual- and group-level) demands. Thus, human behavior in a laboratory setting modeled closely after (7) does not appear to corroborate the simulation data of their agent-based model.[96]

However, we do find evidence that Side Deals among Rich Countries are significantly more binding in "successful" groups—which we define to be those groups who reached agreement without any efficiency losses (i.e. in Rounds 1 and 2)—than in unsuccessful groups. Considering groups who participated in either the PSD or ASD treatments, the Agreed Maximum Demands of the Poor do not significantly differ across successful and unsuccessful groups. However, if we look at groups in either the RSD or ASD treatments, the Agreed Maximum Demand of the Rich significantly differs across successful and unsuccessful groups (Wilcoxon-Mann-Whitney (WMW) test, 62.3 in successful groups vs. 72.6 in unsuccessful groups, p=0.028, z=2.193). In fact, these differences hold—albeit with reduced statistical power—if we compare these groups within RSD (WMW, 58.4 vs. 66.6, p=0.106, z=1.616) and within ASD (WMW, 65.5 vs. 78.6, p=.067, z=1.830) separately. This result indicates that the extent to which high-emission countries tie their hands is of paramount importance for group success, though the same does not hold for low-emission countries.

[96] For further analysis and robustness checks, see part (f) of SI.

Unconditional cooperation. We can also examine the effect of group composition on negotiation success in terms of proportion of individuals inclined to cooperate unconditionally, where "unconditional cooperators" are those who demand at most a percentage equivalent to the Global Target (T=60%) in Round 1. Pooling all treatments together, we find that there is almost exactly one additional unconditional cooperator on average in successful groups, as compared to unsuccessful groups (WMW, 3.89 vs. 2.86, p=0.003, z=-2.945). This result remains intact when we exclude SYM from the comparison (WMW, 3.821 vs. 2.647, p=0.007, z=-2.703).

We further investigate the importance of Rich versus Poor cooperation and find that successful groups have almost double the number of Rich unconditional cooperators as unsuccessful groups, on average (WMW, 1.679 vs. 0.882, p=0.001, z=-3.426), while successful groups and unsuccessful groups are not significantly different in terms of the number of Poor unconditional cooperators (WMW, 2.14 vs. 1.76, p=0.400, z=-0.842). Taken together, these results reinforce the notion that strong commitment and unconditional cooperation by Rich Countries hold paramount influence in determining the success of multilateral negotiations.

ii. Individual demands

Wealth redistribution. An interesting question pertains to the behavior of the two different player types in the asymmetric treatments: is there evidence of redistribution from the Rich to the Poor, in the form of lower demands by the wealthy? In asymmetric groups, we find evidence of such redistribution: the Poor demand 66.7% of initial wealth and the Rich demand 60.2% in the first round (i.e. across all groups in the sample), on average (WMW, p=0.000, z=3.381). More interesting is the apparent dependence of this disparity on whether Side Deals take place prior to the first global negotiation stage. Comparing the average initial demands of Poor and Rich Countries within treatment groups (Figure 2), we see substantial differences under PSD (WMW, 67.3 for Poor vs. 57.8 for Rich, p=0.071, z=1.805), RSD (WMW, 66.4 for Poor vs. 58.3 for Rich, p=0.031, z=2.154), and ASD (WMW, 66.4 for Poor vs. 60.8 for Rich, p=0.092, z=1.686), though this difference is attenuated in ASYM (WMW, 66.7 for Poor vs. 62.9 for Rich, p=0.240, z=1.186). Consistent with (16), it thus appears that Side Deals increase the salience of the inequality, inciting fairness motivations that are manifested through a downward shift in Rich Countries' demands.

This increased salience is especially apparent when the Side Deals pertain to only one subgroup (i.e. *either* the Poor *or* the Rich), as evidenced by the Side Deal inputs (i.e. Maximum Demands) chosen by Poor and Rich negotiators in the various treatments containing Side Deals (see Figure 3). For instance, in PSD, the modal Maximum Demand input in the Side Deal pertaining to the first two rounds of Global Negotiation is 100%, and a vast majority of Poor Countries chooses values at or above the Global Target of 60%. On the contrary, in RSD, not a single player chooses a preferred Maximum Demand of 100%, and a majority of Rich Countries selects a value in the range of 50-70%. However, when both Poor and Rich Countries engage in Side Deals, the distribution of Maximum Demands between the two player types is strikingly similar. Hence, negotiators' decisions are clearly shaped by the initial conditions and institutional frameworks surrounding the bargaining process.

Conditional demands. We additionally explore whether other group members' demands are an important determinant of individual decisions. Indeed, we find evidence of "carbon leakage" across country types; that is, we find a significant positive effect of past cooperation by the Rich (Poor) on Poor (Rich) Countries' demands (Table 3). Specifically, Poor Countries increase their average demand in the present round by almost four percentage points for every additional Rich Country that cooperated (by demanding a percentage less than or equal to the target) in the previous round. Similarly, Rich Countries increase their demands by almost three percentage points for each additional Poor Country that cooperated in the prior round. We do not find evidence that Countries take advantage of the cooperation of like Countries.

4. Discussion

We explore the impact of country heterogeneity and minilateral agreements on climate bargaining processes in a controlled laboratory setting. Our findings stress the importance of early unconditional cooperation by high emitters in efficiently allocating emissions reductions consistent with a global reduction target. However, the experimental data also suggest that some degree of carbon leakage may take place, in the sense that ambitious commitments from high emitters may reduce the abatement efforts of low emitters. That is, we find evidence that the two player types tend to take advantage of the other type's cooperation, demanding to keep a proportion of emissions closer to their BAU as the other type's concessions increase.

We do not find that "tying your hands" ahead of the inclusive negotiations necessarily promotes cooperation, although Side Deals among various subsets of players do affect bargaining dynamics. Importantly, under conditions of heterogeneity, the disparity between the average demands of the two negotiator types widens in the presence of Side Deals, suggesting an even larger role for high-emission (i.e. industrialized or newly industrializing) countries.

What are the implications for international climate negotiations going forward? In light of the vast heterogeneity across countries in terms of both wealth and emissions, the above findings suggest that the infrastructure around which climate change negotiations revolve are crucial determinants of process dynamics. Specifically, our results indicate that low-emission countries will not increase their ambitions in the near term as a result of side agreements by high-emission states, such as the joint announcement made by China and the United States late in 2014. Therefore, high-emission countries will likely need to commit to still further reductions to maintain a current trajectory consistent with limiting mean global temperature rise to 2°C (26). Furthermore, given the strong initial commitments by high emitters necessary to ensure success, the tendency to free ride off of unlike countries means that (generally poor) low-emitting countries—so long as they remain as such—are unlikely to increase their ambitions over time. A prompt and effective agreement thus hinges on strong, unconditional commitments by industrialized and newly industrializing countries, a condition that led to strong contention under the framework of the Kyoto Protocol.

Notwithstanding the recent non-binding global agreement at COP 21 in Paris—which depends on future negotiations to close the gap between INDC pledges and the requisite emissions reductions to keep with the 2°C threshold—the above conclusions cast a shadow on the prospects for a sufficiently ambitious outcome of ongoing global climate negotiations. Our results indicate that minilateral agreements are not "game changers", at least not without significantly ambitious reduction commitments by high-emission countries, which thus far have not materialized. To make matters worse, while the game analyzed here brings potentially disruptive wealth and responsibility heterogeneities to center stage, un-modeled obstacles further hinder climate change cooperation. For instance, the game equates current emissions with responsibilities, neglecting historical accountability and future development requirements. Moreover, only six negotiators must strike an agreement, which simplifies the coordination problem faced by the 197 parties to the UNFCCC.

Importantly, negotiators outside the lab have to rely mostly on voluntary commitments lacking legal force, as demonstrated by the shift from legally binding emission targets to pledge and review mechanisms witnessed in the Paris COP in December 2015. Hence, our voting system for determining the maximum allowable demands in the global negotiations may oversimplify the task of "tying one's hands" compared to the real negotiations, where processes leading to minilateral agreements may vary and countries face incentives to renege on earlier promises if they stand to gain from doing so. However, committed coalitions may use the threat of diplomatic and economic measures, such as "naming and shaming" and trade sanctions, in order to induce cooperation by less ambitious states. Indeed, there are examples of international agreements without binding rules that were successful despite their voluntary nature and reliance on international scrutiny, such as the Helsinki Declaration on human rights (27).

On the other hand, climate negotiations can rely on more instruments than those available to our subjects. Here there are no direct transfer mechanisms, such as the Adaptation Fund and climate finance. In addition, climate co-benefits may lure countries to join small-n clubs early on, providing much needed leadership (1-3). Our game focuses on short-run costs of mitigation, neglecting such opportunities. Yet, policy tends to be defined by short-term incentives and high discounting, as confirmed by the insufficient ambition of the INDCs pledged prior to COP 21 (6, 28). Hence, under the current framework, the global community runs the risk of bargaining toward ineffective agreements in the coming crucial decades. We therefore urge policymakers to consider additional complementary or stand-alone mechanisms to increase the likelihood of avoiding dangerous climate change.

REFERENCES

1. Keohane, R. & Victor, D. The regime complex for climate change. *Persp. on Pol.* **9,** 7-23 (2011).
2. Ostrom, E. Nested externalities and polycentric institutions: must we wait for global solutions to climate change before taking actions at other scales? *Econ. Theory* **49,** 353-369 (2010).
3. Victor, D. Toward effective international cooperation on climate change: numbers, interests and institutions. *Global Environ. Polit.* **6,** 90-103 (2006).
4. Barrett, S. A portfolio system of climate treaties. *Post-Kyoto International Climate Policy: Implementing Architectures for Agreement*, eds. Aldy, J. and Stavins, R. (Cambridge Univ. Press, 2010), pp. 240-270.
5. Falkner, R., Stephan, H. & Vogler J. International climate policy after Copenhagen: towards a 'building blocks' approach. *Global Pol.* **1,** 252-262 (2010).
6. "INDCs as communicated by parties." <http://www4.unfccc.int/submissions/indc/Submission%20Pages/submissions.aspx>.
7. Smead, R., Sandler, R., Forber, P. & Basl, J. A bargaining game analysis of international climate negotiations. *Nat. Clim. Chang.* **4,** 442-445 (2014).
8. Nash, J. The bargaining problem. *Econometrica* **18,** 155-162 (1950).
9. Kalai, E. & Smorodinsky, M. Other solutions to Nash's bargaining problem. *Econometrica* **43,** 513-518 (1975).
10. Muthoo, A. *Bargaining Theory with Applications*. (Cambridge Univ. Press, 1999).
11. Barrett, S. & Dannenberg, A. Climate negotiations under scientific uncertainty. *Proc. Natl Acad. Sci. USA* **109,** 17372-17376 (2012).
12. Barrett, S. & Dannenberg, A. Sensitivity of collective action to uncertainty about climate tipping points. *Nat. Clim. Chang.* **4,** 36-39 (2013).
13. Dannenberg, A., Löschel, A., Paolacci, G., Reif, C. & Tavoni A. On the provision of public goods with probabilistic and ambiguous thresholds. *Environ. Res. Econ.* **61,** 365-383 (2015).
14. Hasson, R., Löfgren, Å. & Visser M. Climate change in a public goods game: investment decision in mitigation versus adaptation. *Ecol. Econ.* **70,** 331-338 (2010).
15. Milinski, M., Sommerfeld, R., Krambeck, H., Reed, F. & Marotzke J. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proc. Natl Acad. Sci. USA*, **105,** 2291-2294 (2008).

16. Tavoni, A., Dannenberg, A., Kallis, G. & Loschel, A. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proc. Natl Acad. Sci. USA* **108,** 11825-11829 (2011).
17. Lenton, T. *et al.* Tipping elements in the Earth's climate system. *Proc. Natl Acad. Sci. USA* **105,** 1786-1793 (2008).
18. Pacheco, J., Vasconcelos, V. & Santos F. Climate change governance, cooperation and self-organization. *Phys. Life Rev.* **11,** 573-586 (2014).
19. Santos, F. & Pacheco, J. Risk of collective failure provides an escape from the tragedy of the commons. *Proc. Natl Acad. Sci. USA* **108,** 10421-10425 (2011).
20. Steffen, W. *et al.* Planetary boundaries: guiding human development on a changing planet. *Science* **347,** 1259855-1259855 (2015).
21. Tavoni, A. Game theory: building up cooperation. *Nat. Clim. Chang.* **3,** 782-783 (2013).
22. Vasconcelos, V., Santos, F. & Pacheco J. A bottom-up institutional approach to cooperative governance of risky commons. *Nat. Clim. Chang.* **3,** 797-801 (2013).
23. Rubinstein, A. Perfect equilibrium in a bargaining model. *Econometrica* **50,** 97-109 (1982).
24. Güth, W., Schmittberger, R. & Schwarze, B. An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Organ.* **3,** 367-388 (1982).
25. Forsythe, R., Horowitz, J., Savin, N. & Sefton, M. Fairness in simple bargaining experiments. *Game. Econ. Behav.* **6,** 347-369 (1994).
26. Friedlingstein, P. *et al.* Persistent growth of CO_2 emissions and implications for reaching climate targets. *Nat. Geosci.* **7,** 709-715 (2014).
27. Bodansky, D. in *Towards a Workable and Effective Climate Regime* (ed, S. Barrett, C. Carraro, and J. de Melo) 155-166 (VoxEU eBook, 2015).
28. International Energy Agency. World energy outlook special report: Executive summary, Paris (2015).

TABLES

Table 1 | Game design

	SYM	ASYM	PSD	RSD	ASD
Endowments	All (×6): £16.67	Poor (×4): £10 Rich (×2): £30	Poor (×4): £10 Rich (×2): £30	Poor (×4): £10 Rich (×2): £30	Poor (×4): £10 Rich (×2): £30
Side Deals	None	None	Poor	Rich	Poor Rich

Table 2 | Success rate by target level

	Rounds 1-2	Rounds 1-4	Rounds 1-6	Rounds 1-8
SYM	63.6%	100.0%	100.0%	100.0%
ASYM	64.3%	78.6%	85.7%	85.7%
PSD	80.0%	80.0%	90.0%	90.0%
RSD	50.0%	70.0%	90.0%	100.0%
ASD	54.5%	72.7%	90.9%	100.0%

Notes: Percentages indicate the proportion of groups in each treatment who had reached agreement by a given threshold round.

Table 3 | Conditional demands of Poor and Rich

	Poor Demand	Rich Demand
Rich Cooperated	3.865**	1.694
	(1.768)	(2.540)
Poor Cooperated	-0.020	2.685***
	(1.047)	(0.813)
Constant	59.401***	53.175***
	(6.194)	(3.578)
Groups	26	26
Subjects	104	52
Obs	356	178
Controls	Yes	Yes

Notes: The dependent variable in this regression indicates the individual demands over the course of negotiation. The independent variables represent the number of Rich and Poor Country representatives (respectively) who cooperated in the prior round by demanding less than or equal to the Global Target. Controls include gender, Annex 1 nationality, stated primary motivation, Global Target, and the difference between the group demand and the target in the prior round of negotiations. There are 26 groups in heterogeneous treatments that negotiated past the first period, and these are the groups considered here. Robust errors are clustered at the group level. Standard errors are reported below estimates in parentheses. ***p<0.001, **p<0.05, *p<0.10

FIGURES

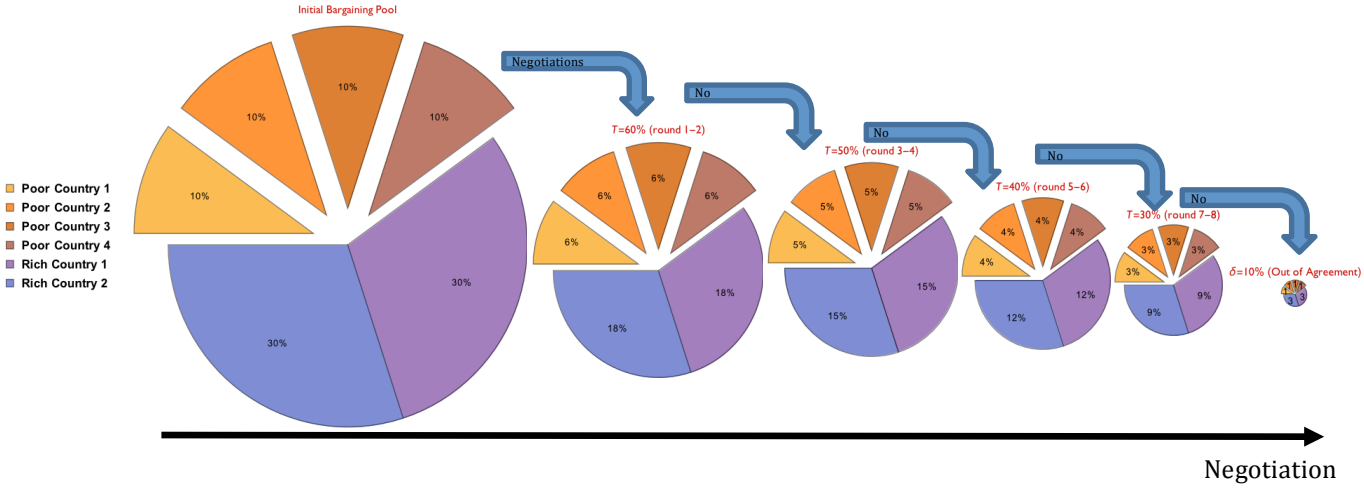


Figure 1 | Timing and dynamics of the game. The six-player bargaining game begins with a collective "pie" of £100, which is split between two Rich Countries (each endowed with 30% of the pie, i.e. £30), and four Poor Countries (each endowed with 10% of the pie, i.e. £10). Starting from this initial allocation of wealth/emissions, the group faces sequential rounds of bargaining on progressively tighter targets. The figure depicts the wealth/emissions distribution ensuing from each target if Countries were to reduce symmetrically.

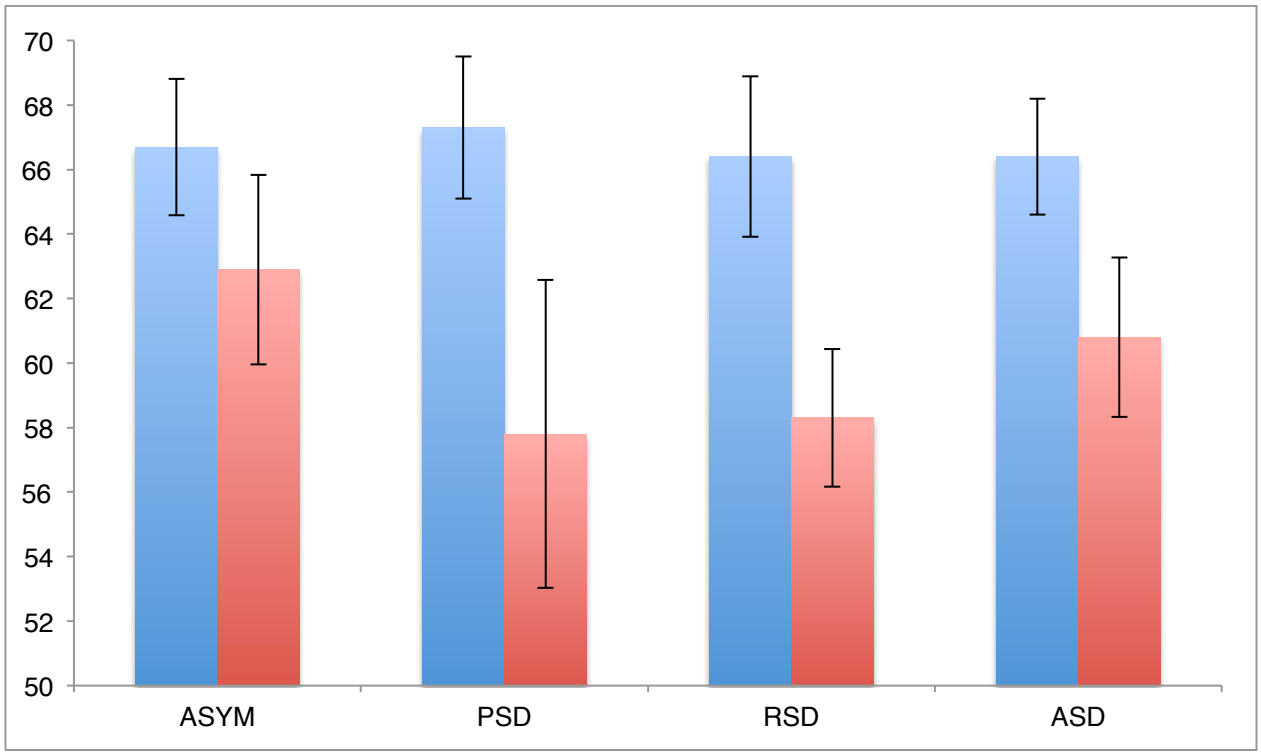


Figure 2 | Average initial demands (and standard error bars) by Poor (blue) and Rich (red) negotiators in treatments with asymmetric endowments.

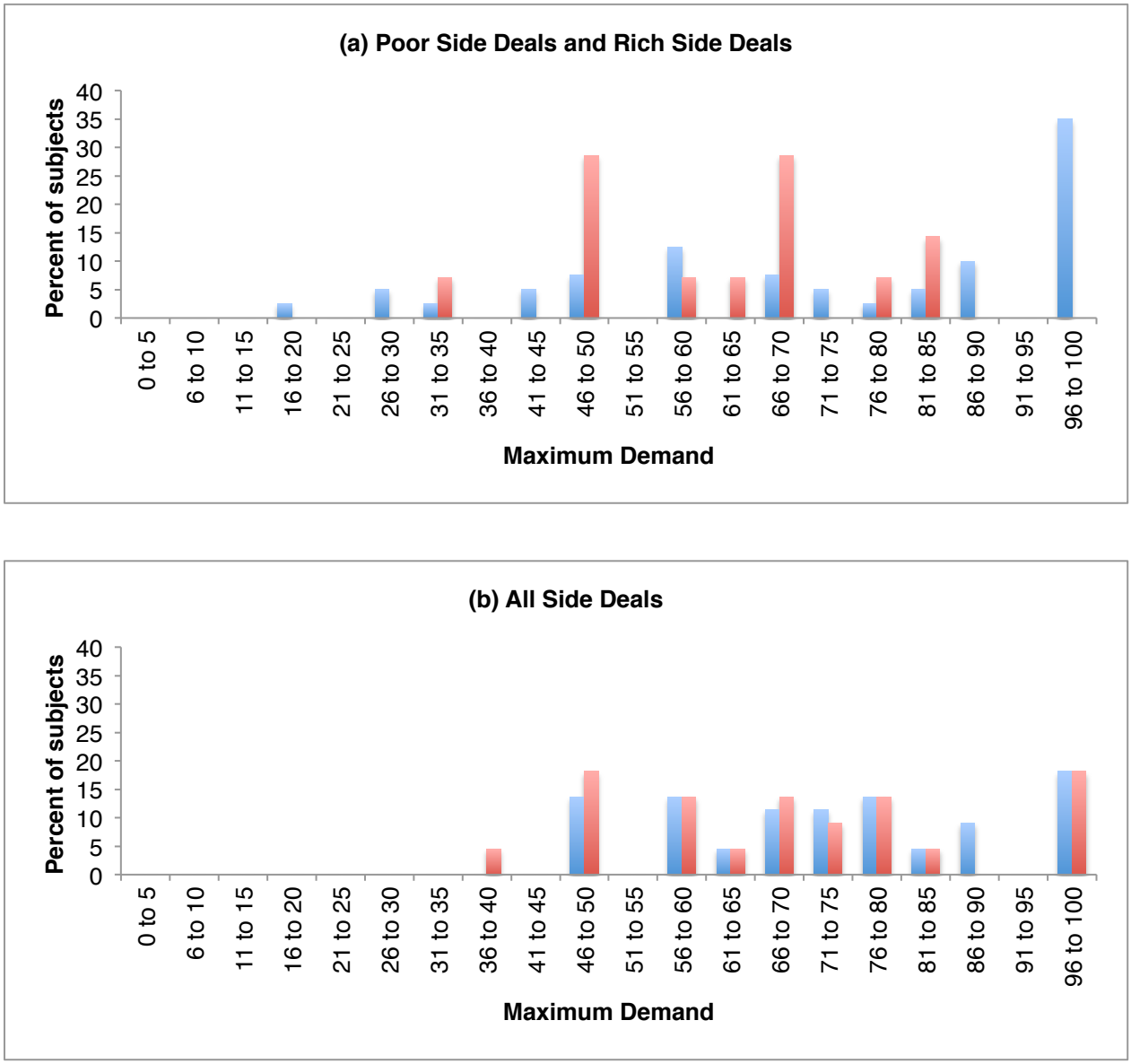


Figure 3 | Distributions of Maximum Demands by Poor (blue) and Rich (red) players in treatments with **(a)** Poor or Rich Side Deals (PSD or RSD), and **(b)** both Poor and Rich Side Deals (ASD). Since only Poor (Rich) Countries input Maximum Demands in the Poor (Rich) Side Deals treatment, (a) combines the data from these two treatments for ease of comparison.

SUPPLEMENTARY INFORMATION

(a) Experimental Design and Related Literature

In an effort to construct an experiment that captured important elements of abrupt climate change yet retained the simplicity necessary to ensure internal validity in a laboratory setting, we made several simplifying design decisions. In the main text we mention these decisions in our discussion of how the game relates to (or departs from) the literature. Here we expand on the motivations for and implications of such choices.

While the experimental literature on climate change negotiations tends to center upon public goods games, we depart from this mechanism in several ways for two primary purposes: a) to enhance the relevance of the context, and b) to provide an empirical test of the agent-based model proposed in (1). Rather than employing a voluntary contribution mechanism devoid of context, we narrow our interest to pertain solely to climate change negotiations, where the instructions provide clear background information on the economic complexities associated with this pervasive externality.

For instance, the dynamic nature of the Global Target captures the cost of delaying legislation to curb greenhouse gas emissions, a stock pollutant with long-term atmospheric warming effects. The target persists over two rounds to allow for learning. While any time lapse theoretically increases the necessity of stronger future abatement commitments to reach a given target, the slow and lagged process of climate change and the relative frequency of negotiations allows for fairly stable global goals in the short term, so that learning can occur from one negotiation to the next. Note that this game neglects gradual damages, since we are concerned with the large costs associated with failure to reach a timely agreement on a target. Such targets can be interpreted as either emerging from scientific evidence or from political discourse (for example, 2ºC).

In addition to the detailed description of global climate change provided in the instructions, various features of the game—e.g., the responsibility dichotomy between asymmetric Country types, the termination of negotiation if the target is reached, and the correlation between emissions and wealth—were designed to mimic the climate context as closely as possible while maintaining the simplicity necessary to ensure the game's comprehensibility. For instance, assigning players to represent either "Rich" (high-emission) or "Poor" (low-emission) Countries mimics the accepted categorization of countries in the COP negotiations, where much diplomatic effort revolves around sharing the "common but differentiated responsibilities" between developed Annex I countries and less developed non-Annex I countries.[97] It additionally captures elements of current emissions responsibility as well as the disproportionate sacrifice associated with deviating from BAU.

Finally, the composition of the groups—where a third of the countries represented are responsible for 60% of global greenhouse gas emissions—is reflective of the 54% for which the top three global players (United States, European Union, and China)—who engaged in pre-COP 21 minilateral discussions—are responsible (2).[98] To ensure that we had sufficient statistical power to detect meaningful differences across treatment groups with and without Side Deals, we did not

[97] "Rich" and "Poor" subjects may behave differently than they would in a symmetric setup, where such labels are not assigned, as indicated in (1). Therefore, we also introduced a symmetric treatment where subjects receive equal endowments, which serves as a baseline towards which wealth heterogeneity can be assessed. We note, however, that division into two groups—regardless of labels—may be considered an additional departure of the ASYM treatment to the SYM treatment. However, an arbitrary division of SYM groups into two groups of unequal sizes may have further confused the instructions, so we opted to ignore this slight departure and assume it holds negligible bearing on the results.

[98] In addition to mimicking real-world heterogeneity, our experiment shares the rich-poor dichotomy with the theoretical investigation in (1)—whose pertinent findings inspired our behavioral investigation—and with related experimental literature pertaining to allocation of emissions in the context of climate change (e.g., 3-6). Our experimental design encapsulates additional features from similar experiments, namely (3). In their experiment, groups composed of six (asymmetrically endowed) players aim to avoid the losses associated with catastrophic climate change in a dynamic framed experiment. However, unlike (3), we do not impose a set number of rounds, and we do not vary the probability of climate catastrophe if the target is met. In our game, meeting the target guarantees payout, but payout is already associated with sacrifices compared to the status quo (i.e. the initial endowment), as in the COP negotiations. Additionally, players in our asymmetric treatments received heterogeneous endowments, whereas those in (3) received symmetric endowments and a subset of players were 'forced' to contribute to a climate fund in the first three rounds to create asymmetry for the following seven rounds.

allow Countries in these treatments to opt out of Side Deals. While in real negotiations it is often possible to avoid public commitments, political pressure may make doing so costly. One could argue, by way of example, that the pressure for China and the U.S. to form an agreement in the run-up to COP 21 together with pending U.S. presidential elections meant that opting out of this "side deal" was not a politically desirable option for either party. Indeed most countries submitted their Intended Nationally Determined Contributions prior to the Paris conference, even if they were not obliged to do so.

(b) Game Equilibria

As shown in the two propositions contained in the SI to (1), the threshold bargaining game employed here features two types of strict Nash equilibria, which can be either *disagreement outcomes* or *feasible solutions.* In disagreement outcomes, all players are unwilling to make sufficient concessions, i.e. the other five players in one's group demand too much for any single player to facilitate agreement by reducing her demand (so that the threshold in a given round is unattainable). In feasible solutions, the threshold is exactly met and everyone is better off than in disagreement, i.e. every player's demand is larger than δ=10% (the amount one can get out of agreement).

For the sake of tractability, let us focus on the treatments without side deals, which allows us to bypass the strategic implications of constraining future decisions. As illustrated below, for our parameters, feasible solutions are preferable equilibria in the sense that they Pareto dominate the disagreement outcomes, yet free-riding incentives pull players toward disagreement. In our game, there are four thresholds corresponding to different pairs of rounds—T=60%, T=50%, T=40%, and T=30%—so essentially one can treat each pair of rounds as a separate game where bargaining takes place on the relevant T. Given the relative values of δ and T in the game, along with the shrinking

of *T* over time, groups maximize their payoffs by coordinating on a feasible solution in the first two rounds. As bargaining continues to later stages, wealth is inevitably lost due to the tightening target and agreement becomes less appealing. Note however, that regardless of the distribution of endowments, players always have an incentive to strike an agreement compared to a disagreement outcome. For example, suppose that the negotiations reached the final round. Failing to strike an agreement would mean a take-home payoff of roughly £1.7 in SYM (and £1 and £3 for Poor and Rich, respectively, in all other treatments). These values are lower than the payoff that players can secure by each demanding to keep 30% of their endowments in Round 8 (i.e. £5 in SYM, £3 and £9 for Poor and Rich, respectively, in all other treatments).

Of course subjects may deviate from symmetric behavior, perhaps due to the presence of obstinate free-riders. While this matter is an empirical one, here we briefly show that some degree of free riding may be sustained in the game, so long as a sufficient number of players is willing to compensate such behavior. Let us restrict attention to SYM, for simplicity. When *T*=60%, up to three free riders can be tolerated in the sense that if the other three players are willing to shoulder (equally) the entire burden of shrinking the pie by 40%, they will still earn £3.3 each, which is more than they would earn out of agreement. By a similar token, up to two free riders are sustainable when *T*=50% or *T*=40%, and only one free-rider can be sustained when *T*=30%.

(c) Experimental Implementation

We employ a design that allows for between-subject and between-group analysis. Each subject participated in a group negotiation of up to eight rounds. Once all groups finished the negotiation, subjects were prompted to complete a brief questionnaire to assess motivation, strategic decision-making, and demographic heterogeneity (see section (d) for the experimental instructions, as well as section (e) for the full questionnaire). Additionally, each subject answered a risk-preference

elicitation question equivalent in structure to the standard question used in (7, 8), with payoffs scaled down to 10% of those used in their experiment. The question asked subjects to select one of five incentive-compatible 50-50 gamble options, where payoffs increase linearly in expected payout and "riskiness" of the gamble, as measured by the standard deviation of the two possible payouts, which ranged from £0.60 to £4.20. The outcome of the gamble was determined individually by a coin toss upon payment for the study.

At the beginning of the experiment, subjects received both written and oral instructions. Each subject must correctly complete a test for understanding before the experiment begins. At the end of the experiment, subjects privately received their experimental earnings in cash, in addition to a £5 show-up fee, totaling £16.80 on average. All experimental decisions were made on a computer screen using the experimental software Z-Tree (9).

A total of 336 student (undergraduate and postgraduate) and non-student subjects volunteered to participate in 20 experimental sessions, most comprising three groups of six subjects (four sessions contained only two groups). The experiment took place at the London School of Economics (LSE), though experimental participation is not restricted to LSE students. In our sample, 50.9% of subjects are female, 42.3% are from Annex I countries (and 52.6% are from countries that engaged in "side deals" prior to COP 21: 5.2% USA, 36.3% EU, and 11.1% China), 47.6% are undergraduate students, and 33.6% are graduate students. The average age of our subjects is 23.5 years (SD=5.99). Student participants come from various disciplines (10.4% Business; 14.9% International Policy, Law, or Government; 8.0% Geography & Environment, 13.1% Economics).

(d) Experimental Instructions for participants of the ASD treatment

Welcome to the experiment! In this experiment, you can earn money. In addition to your earnings from the experiment, you will receive a £5 show-up fee. During the course of the experiment,

please do not talk to other participants. We will now read the experimental instructions aloud. Once we have finished reading, raise your hand if you have questions and we will be with you shortly to answer them. At the end of Part A of the instructions you will find some questions that are meant to ensure that you understand the rules of the experiment. Please answer all questions and signal us by raising your hand when you have finished, so that we may check your answers.

Background: Climate change. Climate change is viewed as a serious global environmental problem. The vast majority of climate scientists expects the global average temperature to rise by 1.1-6.4°C before 2100, where a rise of 2°C is generally considered to be dangerous climate change. There is hardly any disagreement that mankind largely contributes to climate change by emitting greenhouse gases, especially carbon dioxide (CO_2). CO_2 originates from the burning of fossil fuels such as coal, oil, or natural gas in industrial processes and energy production, as well as from combustion engines of cars and lorries. CO_2 is a *global* pollutant—that is, each unit of CO_2 emitted has the same effect on the climate regardless of the location where the emissions occur. Dangerous climate change will result in significant global costs, which get worse over time if agreement is not reached. International climate change negotiations involve yearly meetings where delegations representing different *countries try to strike a global agreement on emissions reductions* that are consistent with the goal of avoiding dangerous climate change. Here you will be asked to negotiate such costly emissions reductions on behalf of the Country to which you will be assigned. Your choices, together with those of the other 'Countries', will determine your payout from the experiment.

Rules of play. Now we will introduce you to a game simulating international climate change negotiations. In total, six Countries are involved in the global negotiation. That is, in addition to you, there are five other negotiators in your negotiation group, and each of you represents one Country. The six Countries account for *all* global wealth and CO_2 emissions (for simplicity, we

disregard other greenhouse gases in the experiment). While excessive emissions impose global costs, individual Countries rely on productive processes which create emissions in order to generate wealth: for every 1 billion tons of CO_2 'emitted' in the game, you receive £1. Hence, reducing emissions is costly. Your decisions in the experiment are anonymous. To guarantee anonymity, you will be randomly assigned to one type of Country (Rich or Poor), and you will be identified by one of the following names: Rich Country 1, Rich Country 2, Poor Country 1, Poor Country 2, Poor Country 3, Poor Country 4. Your name will appear on the lower left side of your screen once the experiment begins. At the beginning of the experiment, you will receive a sum of money that represents your Country's wealth. This wealth mirrors your Country's CO_2 emissions. Therefore, throughout the instructions and the experiment, we will refer to wealth and emissions interchangeably. The current situation in your negotiation group can be summarised as follows:

- **Two Rich Countries** each emit 30 billion tons of CO_2 and earn **£30** in doing so;
- **Four Poor Countries** each emit 10 billion tons of CO_2 and earn **£10** in doing so;
- The resulting Global Emissions amount to **100 billion tons of CO_2** (2×30 billion tons of CO_2 + 4×10 billion tons of CO_2)
- Hence, **Global Wealth** is equal to **£100** (2×£30 + 4×£10)

Due to the threat of dangerous climate change, the goal is to agree on an aggregate level of Global Emissions that does not exceed a given Global Target. In the following experiment, you will participate in up to 8 rounds of climate change negotiations, where the global costs from *not* reaching agreement increase every 2 rounds. **Accordingly, the Global Target decreases every two rounds, as follows:**

- Rounds 1-2: **60% of current emissions** (60 billion tons of CO_2)
- Rounds 3-4: **50% of current emissions** (50 billion tons of CO_2)

- Rounds 5-6: **40% of current emissions** (40 billion tons of CO_2)
- Rounds 7-8: **30% of current emissions** (30 billion tons of CO_2)

To be clear, since current global emissions are 100 billion tons of CO_2, an agreement is only reached if total negotiated emissions are at most 60 billion tons of CO_2 in the first two rounds. Equivalently, Global Wealth must be reduced from an initial level of £100 to a target level of £60 if the Global Target is to be met in the first two rounds. This target becomes more difficult to meet as the negotiations move forward, as outlined above. Every Country faces a similar decision-making problem. In each round of the global negotiation, all six Countries will be asked simultaneously: *"What percent of YOUR COUNTRY's emissions/wealth do you demand to keep?"* If the required Global Target is met, then your group has reached an agreement; negotiations terminate and each Country receives its demand from that round. If agreement is not reached, the negotiation continues to the next round. If an agreement is not reached by the end of the 8th Round of negotiations, dangerous climate change becomes unavoidable and economic costs for all Countries ensue. Each Country will then receive **10% of its initial wealth** (£3 for Rich Countries, £1 for Poor Countries).

Example 1. Imagine that you are part of a negotiation group that makes decisions as follows. In **Round 1** (Global Target=60%), all Countries demand to keep 90% of their emissions/wealth. If the Global Target were to be met, Rich Countries would receive £27 in payout and Poor Countries would receive £9 in payout. See **Screenshot 1** below, for the screen that will be seen by Poor Country 1. However, the Global Target is NOT met and negotiations continue to Round 2. In **Round 2** (Global Target=60%), demands are as follows:

- Rich Country 1 and Poor Country 1 each demand to keep 50%. If the Global Target were to be met, Rich Country 1 would receive 50% of its initial wealth (£15) and Poor Country 1 would receive 50% of its initial wealth (£5).
- Rich Country 2 and all remaining Poor Countries (2,3,4) each demand to keep 80%. If the Global Target were to be met, Rich Country 2 would receive 80% of its initial wealth (£24) and Poor Countries 2, 3, and 4 would receive 80% of their initial wealth (£8 each).

See **Screenshot 2** below. However, Global Demand=68% > Global Target = 60%, so the Global Target is not met and negotiations continue. *Now imagine that the negotiation group continues to demand to keep emissions/wealth above the target level until the 7th Round, when the relevant Global Target is 30% of emissions/wealth.* In **Round 7**, demands are as follows:

- Rich Country 1 and Poor Country 4 demand to keep 32% each.
- Rich Country 2 and Poor Countries 1, 2, and 3 demand to keep 20% each.

See **Screenshot 3**.

Hence, Global Demand = 25% ≤ Global Target = 30%. The Global Target is met. Rich Country 1 receives 32% of its initial wealth (£9.60), Rich Country 2 receives 20% of its initial wealth (£6), Poor Countries 1, 2, and 3 each receive 20% of their initial wealth (£2 each), and Poor Country 4 receives 32% of its initial wealth (£3.20). Please take a brief moment to review and understand the rules, then continue to the next page to test your understanding.

Control questions. **Test your understanding:** *For the questions below, please check the box of the correct answer or fill in your answer on the line provided. For convenience, we summarised the main rules below:*

Global Target
Rounds 1-2: **60%**

Rounds 3-4: **50%**
Rounds 5-6: **40%**
Rounds 7-8: **30%**

Country Initial Wealth
Rich Country 1, Rich Country 2: **£30**
Poor Country 1, Poor Country 2, Poor Country 3, Poor Country 4: **£10**

1. In Round 4's global negotiation, all members of your negotiation group demand to keep 60% of their initial emissions/wealth. What happens next?

- ☐ *We've met our Global Target; each of us receives 60% of our initial wealth.*
- ☐ *Our Global Target has not been met; we continue to Round 5.*

2. In Round 3's global negotiation, all Rich Countries demand to keep 50% of their original emissions/wealth. If two Poor Countries demand to keep 40% and the other two Poor Countries demand to keep 60%, is agreement reached?

- ☐ Yes
- ☐ No

If yes, how much does each Country receive (without show-up fee)? If no, please leave blank.

Rich Countries: £_____each

Poor Countries that demanded 60%: £_____each

Poor Countries that demanded 40%: £_____each

3. In the final Round's global negotiation (i.e. Round 8), one Rich Country demands to keep 20% of its initial emissions/wealth and the other Rich Country demands to keep 30%. If two

Poor Countries demand to keep 30% each and the other two Poor Countries demand to keep 75% each, is agreement reached?

- ☐ Yes
- ☐ No

How much does each Country receive as their final payout (without show-up fee)?

Rich Country that demanded 20%: £_____

Rich Country that demanded 30%: £_____

Poor Countries that demanded 30%: £_____each

Poor Countries that demanded 75%: £_____each

Please raise your hand when you have answered all questions, and we will come to check your answers.

Side Deals. Recall that the Global Target changes every two rounds. Before global negotiations on a new target begin, both groups of Countries (the 4 Poor and the 2 Rich) will simultaneously enter into separate Side Deals, as follows.

(i) Side Deal for Poor Countries:

Prior to the global negotiations in Rounds 1, 3, 5, and 7, each Poor Country will enter its preferred 'Maximum Demand', i.e. the desired maximum percentage of emissions/wealth that each *Poor Country* may demand to keep in the two upcoming global negotiations.

The average of these four Maximum Demands will determine the 'Agreed Maximum Demand for Poor', which cannot be exceeded by each Poor Country in the two upcoming global negotiations.

(ii) Side Deal for Rich Countries:

At the same time, and prior to the global negotiations in Rounds 1, 3, 5, and 7, each Rich Country will enter its preferred 'Maximum Demand', i.e. the desired maximum percentage of emissions/wealth that each *Rich Country* may demand to keep in the two upcoming global negotiations.

The average of these two Maximum Demands will determine the 'Agreed Maximum Demand for Rich', which cannot be exceeded by each Rich Country in the two upcoming global negotiations.

Should a global agreement *not* be reached within the first two rounds, a new target will apply to Round 3 (Global Target=50%) and a new Agreed Maximum Demand will be determined by both Poor and Rich Countries for the two upcoming rounds (Rounds 3 and 4). This process will continue until Round 8 so long as a global agreement is not reached. Please refer to the timeline in **Screenshot 4** for a recap on the various stages of the game.

Example 2. Imagine that you are Poor Country 1 and that you have entered into a Side Deal with the other Poor Countries. In the experiment you will see the following screen (**Screenshot 5**).
The choices from the Side Deal for Poor Countries are shown at the top of **Screenshot 6**, which we have highlighted with a box:

- Poor Country 1 (you) chooses Maximum Demand = 100%
- Poor Country 2 chooses Maximum Demand = 66%
- Poor Country 3 chooses Maximum Demand = 33%
- Poor Country 4 chooses Maximum Demand = 0%

The resulting agreed Side Deal is that each Poor Country cannot exceed 50% demand in the two upcoming global negotiations, i.e. the Agreed Maximum Demand = 50%. (Note that the outcomes

of the Side Deal for Rich Countries, which took place at the same time, are also shown in **Screenshot 6**. All Countries see these outcomes.)

Example 3. Imagine that you are Rich Country 1 and that you have entered into a Side Deal with Rich Country 2. In the experiment you will see the following screen (**Screenshot 7**).

The choices from the Side Deal for Rich Countries are shown at the bottom of **Screenshot 8**, which we have highlighted with a box:

- Rich Country 1 (you) chooses Maximum Demand = 75%
- Rich Country 2 chooses Maximum Demand = 25%

The resulting agreed Side Deal is that each Rich Country cannot exceed 50% demand in the two upcoming global negotiations, i.e. the Agreed Maximum Demand = 50%.

(e) Questionnaire

Question	Response
Was the experiment difficult to understand?	Not at all difficult Somewhat difficult Difficult Very difficult Extremely difficult
Please select the MOST important reason for your decisions during the experiment. *Note: the questionnaire also asked for the second and third most important reasons.*	Monetary self-interest Fairness consideration Maximise group performance (i.e. efficiency) Minimise time spent negotiating Beliefs about actual (international/climate) negotiations Past behaviour of group members Other
If you could redo the experiment, how would you change your choices (if at all)?	Open-ended
In the scenario where 'Rich Country 1' and 'Rich Country 2' are each endowed with 30 billion metric tons in CO2 emissions (or £30) and the four Poor Countres are each endowed with 10 billion tons in CO2 emissions (or £10) each, what do you think would have been a fair initial demand (%) for each of the Rich Countries?	Open-ended (number)
What do you think would have been a fair initial demand (%) for each of the Poor Countries?	Open-ended (number)
Imagine you are in the final round of negotiation. All of the other countries in your group have made their demands and your demand could be pivotal (i.e. 'tip the scale' in terms of whether an agreement is reached or not). In this situation, what is the minimum demand (%) you would accept if you knew that your decision would change the group outcome from non-agreement to agreement?	Open-ended (number)
Now you will select from among five different gambles the one gamble you would like to play. The five different gambles are listed below. You must select one and only one of these gambles. Each gamble has two possible outcomes (Event A or Event B), each with a 50% probability of occurring. Your compensation for this part of the study will be determined by: 1) which of the five gambles you select; and 2) which of the two possible events occur.	Gamble 1: £1.00 vs. £1.00 Gamble 2: £1.80 vs. ££0.60 Gamble 3: £2.60 vs £0.20 Gamble 4: £3.40 vs. -£0.20 Gamble 5: £4.20 vs. -£0.60

Please note that if you should select either gamble 4 or gamble 5 and Event B occurs, your losses will be deducted from your show-up fee. For example: If you select gamble 4 and Event A occurs, you will be paid $£3.40. If Event B occurs, you will have £0.20 deducted from your £5 show-up fee. For every gamble, each event has a 50% chance of occurring. At the end of the study, a volunteer will be asked to flip a coin to determine whether Event A (heads) or Event B (tails) will pay out. Please select your preferred gamble and then WRITE THE NUMBER OF THE GAMBLE YOU SELECTED ON YOUR PAYMENT SLIP.	
Are you generally a person who is fully prepared to take risks (risk prone) or do you try to avoid taking risks (risk averse)? Please select from the following options, where 0 means EXTREMELY RISK AVERSE and 10 means EXTREMELY RISK PRONE.	0 1 2 3 4 5 6 7 8 9 10
Have you ever donated money or goods to a charitable organisation? If yes, how frequently?	Very often Often Sometimes Rarely Never
Is global climate change a serious problem?	Extremely serious Very serious Serious Somewhat serious Not at all serious
Which of the following guiding principles describes your understanding of fairness best in the context of international climate negotiations?	a) Countries with high emissions in the past should reduce more emissions. b) Countries with high economic performance should reduce more emissions. c) Countries should reduce their emissions in such a way that emissions per capita are the same for all countries. d) Countries should reduce their emissions in such a way that the emissions percentage is the same for all countries.
How often do you recycle?	Very Often Often Sometimes Rarely Never

Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people? Please tick a box on the scale, where the value 1 means "need to be very careful" and the value 10 means "most people can be trusted". You can use the values in between to make your estimate.	0 1 2 3 4 5 6 7 8 9 10
Finally, how good are you at working with fractions (e.g. "one fifth of something") or percentages ("e.g. "20% of something")?	Extremely good Very good Good Somewhat good Not good at all

(f) Additional Empirical Analysis

Velocity, Dynamics, and Distributions. In terms of agreement velocity, the most successful treatment group is the one allowing for Side Deals among the Poor (PSD), where on average the groups coordinated on the threshold shortly after the second round (Supplementary Table 1). By contrast, RSD is the treatment where agreement is most delayed (3.5 rounds on average). While ASYM and ASD are comparable in terms of the average agreement round, we note that there are two ASYM groups that failed to reach agreement altogether, consistent with the higher variance in agreement round for ASYM than for ASD. Similarly, while SYM and PSD are comparable along the former dimension, one PSD group was not successful in coordinating on the threshold, consistent with the higher variance in outcomes for PSD than for SYM.

As discussed in the manuscript, all symmetrically endowed (SYM) groups maintained at least 50% of the initial pie, which is remarkably efficient given that the maximum attainable proportion of global wealth is 60%. However, the PSD treatment is the most successful in securing agreement under maximally efficient conditions (i.e. in Rounds 1-2, before the target shrinks), though we do not have the power to detect a statistically significant difference between groups' success rates within the first two periods. Regardless, in accordance with (1), endogenous demand restrictions

(i.e. binding Side Deals) on a larger number of low-emission "poor" players appear to be more successful in inducing coordination than similar restrictions on a smaller number of high-emission "rich" players if we are concerned with maximizing the global pie. Importantly, unlike (1), we do not find conclusive evidence that outcomes in treatments containing Side Deals improve upon global negotiations that occur among asymmetric actors in the absence of Side Deals, in terms of either agreement velocity or demands (at both the individual and the group levels).

Supplementary Figures 1 and 2 provide visual representations of the above statistics in addition to the demand dynamics across treatments. The early disparity in agreement rate across treatments is clear, as is the tendency of average group demands to respond to the declining values of the Global Target *T* (from 60% to 30%) by clustering, although with some variance, around these values.

Across successful asymmetric groups, the average difference between Rich and Poor demands in the successful round of negotiation is 10.65 percentage points ($p<0.01$). This average demand distribution translates to a final average income of £15.63 for Rich players and £6.28 for Poor players, and a final wealth distribution of 27.7% of global wealth for Rich Countries and 11.1% of global wealth for Poor Countries. Note that this subtle shift in the wealth distribution occurs solely in the negotiation over emissions reductions (i.e. it is independent of international wealth and technology transfers pervasive in climate change negotiations).

Moreover, in PSD and RSD, the standard error among players in the group who formed the Side Deal (2.45 in PSD, and 3.86 in RSD) is smaller than it is for the subgroup without constraints (6.27 in PSD, and 4.28 in RSD; Supplementary Figure 3). Therefore, in the case of the side agreements among the US, the EU, and China, we would expect low-emission countries to experience a wider variance in proposed emissions limits.

Questionnaire Analysis. Immediately following the experiment, subjects were asked a series of questions to gather demographic information, preferences (i.e. for fairness, risk, environment), and motivations in the experiment.

We look at players' primary decision-making motivations, acknowledging that the *ex post* nature of the questionnaire may create dependence of answers on dynamics and outcomes of the game played previously. When asked what is the most important motivation behind their decisions in the game, most claimed to have been primarily motivated by group efficiency (36.3% of subjects) or money (29.1% of subjects), with fairness (19.6% of subjects) following close behind. The rest were motivated by time minimization (7.5% of subjects), beliefs about actual climate negotiations (3.6% of subjects), and the past behavior of group members (3.6% of subjects). If money was a subjects' primary motivation, she initially demanded 6.9 percentage points more than if her primary motivation were not money (p=0.001).

We do not find that stating fairness as one's primary motivation influences one's initial demand in the SYM treatment. However, when we introduce asymmetric endowments, fairness influences demands considerably. Poor players who stated fairness as their primary motivation (22% of Poor) demanded about 4.5 percentage points more in Round 1 than those who stated another motivation (70.1 percent vs. 65.8 percent, p=0.094), consistent with the self-serving notion of fairness found in ultimatum games with asymmetric payoffs (10). Likewise, Rich players who stated fairness as their primary motivation (16% of Rich) demanded almost 10 percentage points less than those with other motivations (61.7 percent vs. 52.0 percent, p=0.025), consistent with social welfare preferences (11).

Additionally, we asked subjects what is the minimum demand they would accept if they were a pivotal player in the final round of negotiation, i.e. when the Global Target is 30%. The average

minimum acceptable demand is 30.2 percent (SD=16.8) of one's endowment, and this is not largely dependent on whether one was a Poor (mean=31.9, SD=17.2) or Rich (mean=28.5, SD=17.1) player.

We also ask a series of questions to elicit our subjects' risk and social preferences. Using a variant of the incentive-compatible risk preference elicitation question posed in (7, 8)—where 1 represents a certain outcome (50% chance of £1 vs. 50% chance of £1) and 5 represents the most risky outcome (50% chance of £4.20 and 50% chance of -£0.60)— subjects' average selection is 3.77 (SD=1.33). When asked to self-assess the extent to which they are risk prone on a scale from 0 to 10 (where 10 is extremely risk prone; see 12), subjects' average selection is 5.38 (SD=2.11). To assess subjects' altruism, we ask them to state the frequency with which they donate to charity: 6.9% of subjects give to charity *very often*, 17.7% give to charity *often*, 45.8% give to charity *sometimes*, 28.1% give to charity *rarely*, and 1.6% *never* give. We also asked subjects the extent to which they think others can be trusted on a scale from 1 (low trust) to 10 (high trust), and the mean response is 4.9 (SD=2.1). To get a reading of subjects' preferences for the environment, we asked how often the subjects recycle. In our pool, 27.4% claim to recycle *very often*, 39.0% recycle *often*, 19.1% recycle *sometimes*, 4.8% recycle *rarely*, and 9.8% *never* recycle. Additionally, when asked their opinion on the severity of the problem of climate change, 35.0% of subjects responded that it is *extremely serious*, 36.0% that it is *very serious*, 19.6% that it is *serious*, 8.2% that it is *somewhat serious*, and 1.3% that it is *not at all serious*. Group-level heterogeneity in self-reported charitable spending and 'green' preferences do not play a significant role in determining subjects' decision-making nor the velocity of agreement in the game, contrary to the assertion that heterogeneity of preferences increases the transaction costs associated with (and therefore decreases the likelihood of) reaching agreement (13).

To gauge whether subjects understood the experiment, we ask the extent to which the instructions are comprehensible and ask for an evaluation of subjects' own ability to work with fractions. Subjects appear to have understood the experiment, with only six subjects (i.e. less than two percent) stating that the experiment was (very) difficult to understand. Similarly, only 2.3% of subjects claim they are *not at all good* with fractions, while 9.2% are *somewhat good* with fractions, 27.8% are *good* with fractions, 37.3% are *very good* with fractions, and 23.5% are *extremely good* with fractions.

Risk Preferences. To further understand the dynamics underpinning group coordination, we investigate the role of individual risk preferences in predicting behavior in the negotiation. As expected, we find that risk aversion reduces demand, on average. In Supplementary Table 2, we display the effects of risk preferences on individual demand in a panel OLS regression. *Risk* is measured on a scale from 1 to 5, where 1 represents the most risk-averse gamble option—a gamble with payout certainty—and 5 represents the most risk-seeking option.

Supplementary Table 2 demonstrates that the effect of the risk parameter on demand is robust with respect to inclusion of various controls. The initial inclusion of controls—including demographics, stated motivation, Global Target, and treatment group assignment—reduces the magnitude of the effect from 1.68 to 1.24 percentage points per one-point increase on the risk scale. However, the magnitude of the effect is restored when we additionally account for the role of threshold (even) rounds—or rounds in which a failure to reach agreement results in negative group-level consequences— which have a large negative effect on demand, as expected.

We further investigate the role of threshold effects through the interaction term between threshold rounds and risk preferences. Since threshold rounds threaten to diminish global welfare, we expect risk-averse individuals to err on the side of caution by demanding less than risk-prone individuals

in these rounds. In regression four of Supplementary Table 2, we see that the state of being in a threshold round reduces individual demand by almost six percentage points on average. However, the positive coefficient for the interaction term—which is significant both here (p=0.068) and when using self-reported risk attitudes as the independent variable (p=0.036; see Supplementary Table 3)—indicates that this threshold effect is less strongly negative the more risk seeking is the individual.

While risk preferences are important predictors in the game, the question remains as to the interpretation of such results when considering actions taken by countries. We suffice to say here that risk preferences may potentially be an important and understudied predictor of (climate) bargaining strategies, whether they are risk preferences of the negotiators themselves or broader cultural parameters. For example, countries may signal risk attitudes through policies or military and geopolitical strategies, in turn providing information on their bargaining strategies. Our results indicate that risk preferences in bargaining may be a promising avenue for future research.

Self-Serving Bias. Our data allows for empirical estimation of self-serving bias (14, 15). In the questionnaire described above, subjects were asked a series of survey questions, one of which pertained to their perspectives on equity in the context of climate change. To test for self-serving bias, we look at the average marginal effects of logit regressions where the dependent variables are dummies for whether the particular equity perspective in question has been selected, and the independent variables are indicators for subjects' nationality (United States, European Union, or China). Controlling for whether subjects played the role of a Rich or Poor Country in the experiment, we find that European subjects were less likely to identify with the perspective that "Countries with high emissions in the past should reduce more emissions" by 12.95% (p=0.038), although they were somewhat more likely than non-Europeans to identify with the perspective that

"Countries with high economic performance should reduce more emissions" by 9.7% (p=.123). Additionally, we find that Chinese subjects were less likely to select "Countries with high economic performance should reduce more emissions" by 15.64% (p=0.055). We do not find definitive evidence of self-serving bias among Americans in our sample; however, American subjects were less likely to identify with the perspective that "Countries should reduce their emissions in such a way that emissions per capita are the same for all countries" than non-Americans by 13.8%, though the effect is not quite significant at conventional levels (p=0.140).

Supporting Analysis: Robustness. To account for the maximum demand imposed in the experimental design, we run a panel Tobit regression (see Supplementary Tables 4 and 5) to complement the panel OLS regressions previously discussed (see Table 3 in the manuscript and Supplementary Table 2). These regressions place an upper limit of 100 on individual demands. Since subjects may wish to demand more than 100 percent of their endowed share of global emissions, the Tobit regressions censor the dependent variable from above at 100. Note that it is not necessary to censor the dependent variable from below since none of the experimental subjects demanded zero emissions in the game. The results of the Tobit regressions align closely with those of the OLS regressions, providing a simple robustness check of the conditional demand result and the influence of risk preferences on individual demands.

We run an additional panel OLS regression (see Supplementary Table 3), replacing the incentive-compatible risk preference with a stated preference for risk as our dependent variable. Again, the results are qualitatively similar to those in Supplementary Table 2. While the incentive-compatible risk responses map preferences on a scale from 1 to 5, the stated risk responses map preferences on a scale from 0 to 10. Standard errors are slightly inflated relative to the OLS regression on the incentive-compatible risk preference. However, the results for Regressions 1-3 across the two

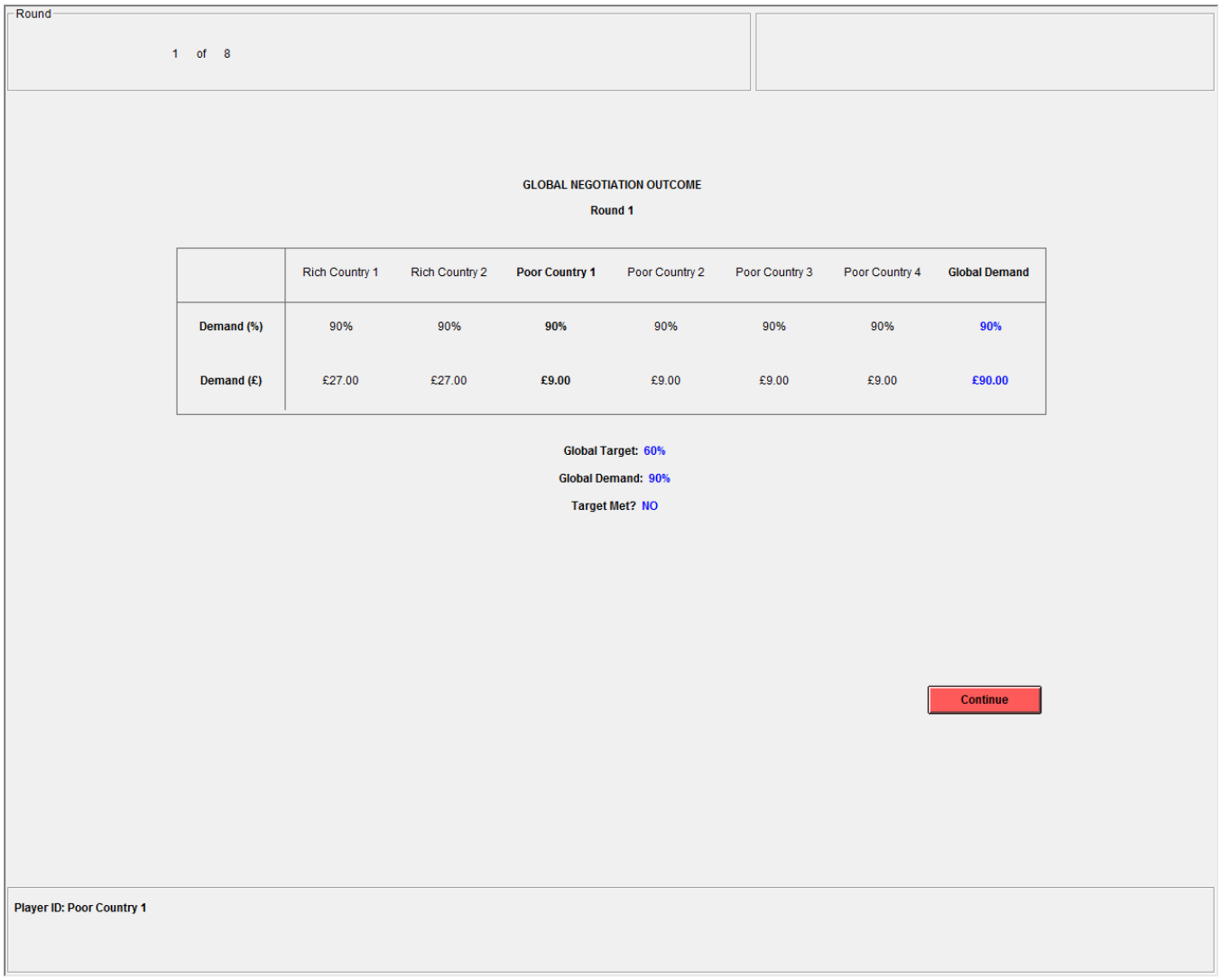
tables are qualitatively similar. Interestingly, the results for Regression 4 indicate a positive though non-significant effect of risk preference on demand in the game, while the interaction between threshold round and risk becomes significant. That is, subjects who indicate a higher risk tolerance demand more in threshold rounds than do those who report lower risk tolerance.

References

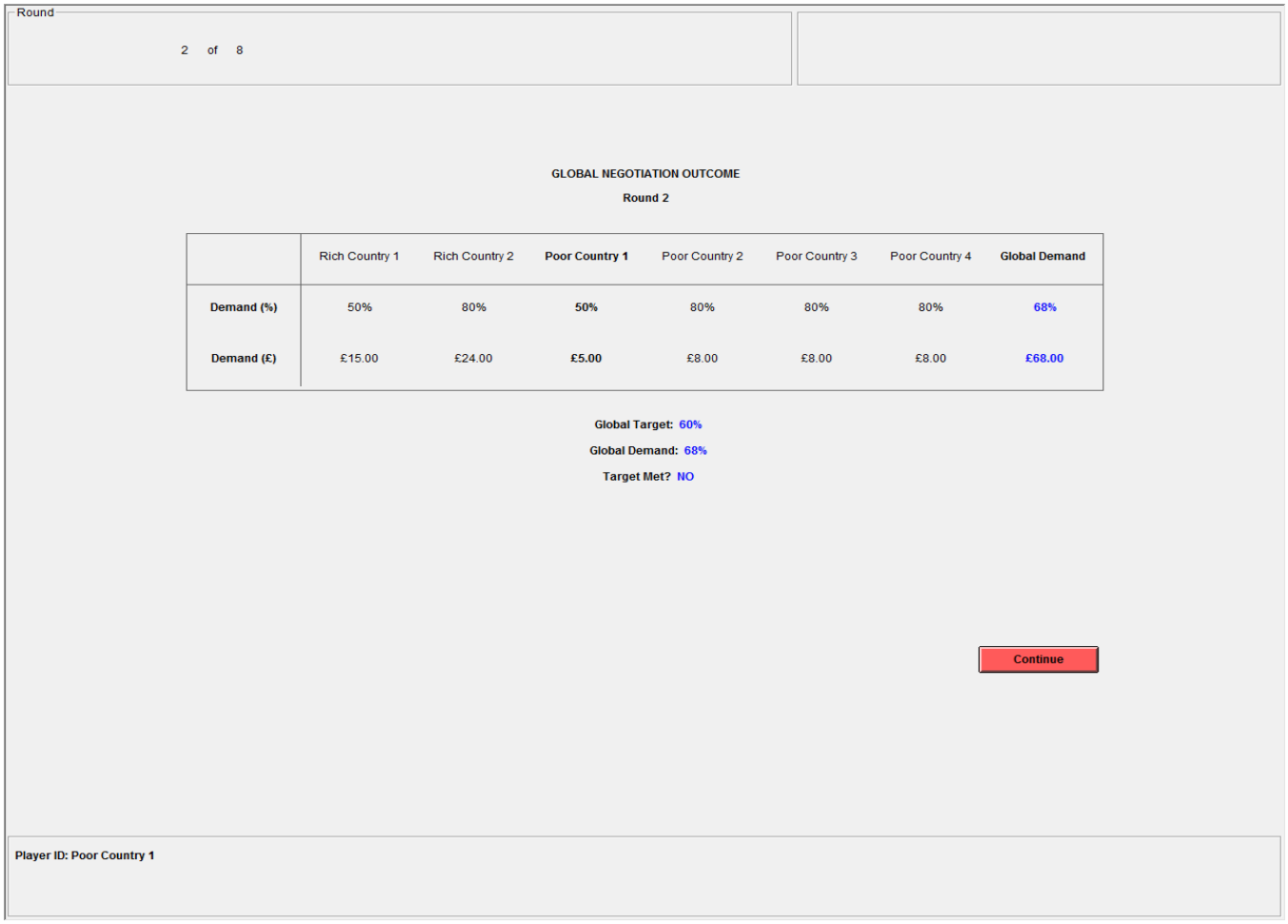
1. Smead R., Sandler R., Forber P., Basl J. A bargaining game analysis of international climate negotiations. *Nat. Clim. Chang.* **4**, 442-445 (2014).
2. Friedlingstein, P. *et al.* Persistent growth of CO_2 emissions and implications for reaching climate targets. *Nat. Geosci.* **7**, 709-715 (2014).
3. Tavoni, A., Dannenberg, A., Kallis, G., & Loschel, A. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proc. Natl Acad. Sci. USA* **108**, 11825-11829 (2011).
4. Milinski, M., Sommerfeld, R., Krambeck, H., Reed, F. & Marotzke J. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proc. Natl Acad. Sci. USA*, **105,** 2291-2294 (2008).
5. Burton-Chellew, M., May, R., West, S. Combined inequality in wealth and risk leads to disaster in the climate change game. *Clim. Chang.***120**, 815-830 (2013).
6. Gampfer, R. Do individuals care about fairness in burden sharing for climate change mitigation? Evidence from a lab experiment. *Clim. Chang.* **124**, 65-77 (2014).
7. Eckel, C. & Grossman, P. Forecasting risk attitudes: an experimental study using actual and forecast gamble choices. *J. Econ. Behav. Organ.* **68,** 1-17 (2008).

8. Binswanger, H. Attitudes toward risk: experimental measurement in rural India. *Am. J. Agr. Econ.* **62**, 395-407 (1980).
9. Fischbacher, U. z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* **10,** 171-178 (2007).
10. Kagel, J., Kim, C, & Moser, D. Fairness in ultimatum games with asymmetric information and asymmetric payoffs. *Game. Econ. Behav.* **13**, 100-110 (1996).
11. Charness, G. & Rabin, M. Understanding social preferences with simple tests. *Q. J. Econ.* **117**, 817-869 (2002).
12. Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. Individual risk attitudes: Measurement, determinants, and behavioural consequences. *J. Eur. Econ. Assoc.* **9**, 522-550 (2011).
13. Libecap, G. Addressing global environmental externalities: transaction costs considerations. *J. Econ. Lit.* **52**, 424-479 (2014).
14. Lange A. & Vogt C. Cooperation in international environmental negotiations due to a preference for equity. *J. Pub. E.* **87**, 2049-2067 (2003).
15. Lange, A., Löschel, A., Vogt, C., & Ziegler, A. On the self-interested use of equity in international climate negotiations. *Eur. Econ. Rev.* **54**, 359-375 (2010).

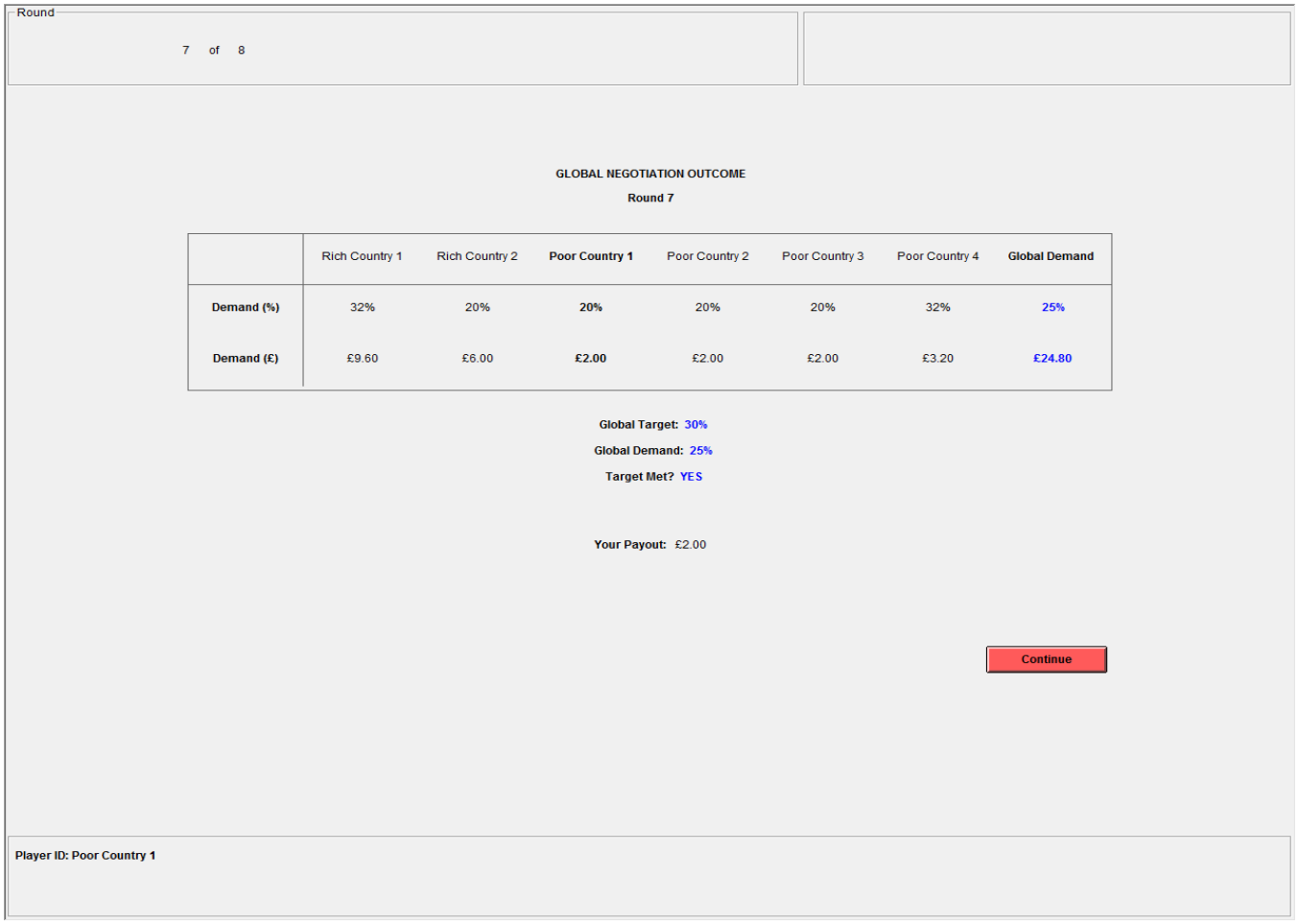
Screenshots from game interface (from the experimental instructions)



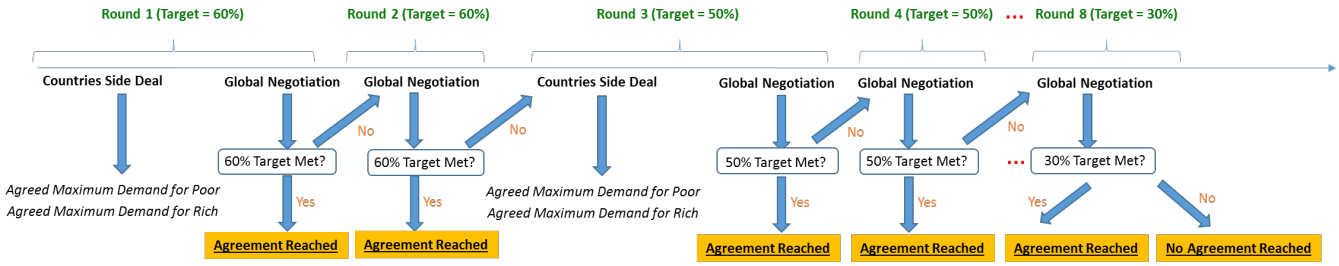
Screenshot 1 I Outcome screen presented to Poor Country 1 in Round 1 of the Global Negotiation if all group members demand to keep 90% of their initial wealth/emissions in ASYM, PSD, RSD, and ASD. The Global Demand exceeds the Global Target of 60% in Round 1 and negotiations continue to Round 2.



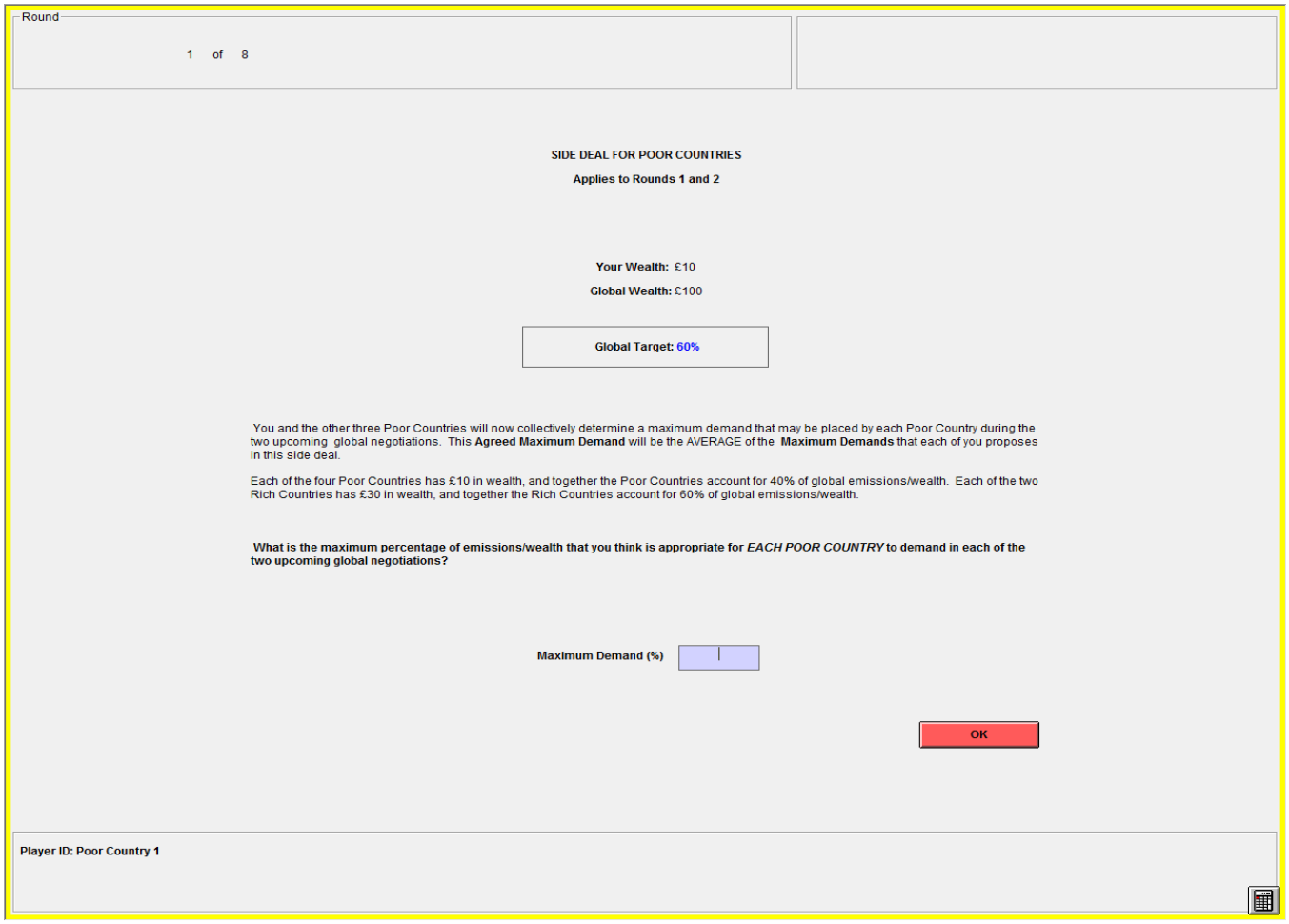
Screenshot 2 I Outcome screen presented to Poor Country 1 in Round 2 of the Global Negotiation if Rich Country 1 and Poor Country 1 demand to keep 50% of their initial wealth/emissions and all other players demand to keep 80% of their initial wealth/emissions. The Global Demand exceeds the Global Target of 60% in Round 1 and negotiations continue to Round 3.



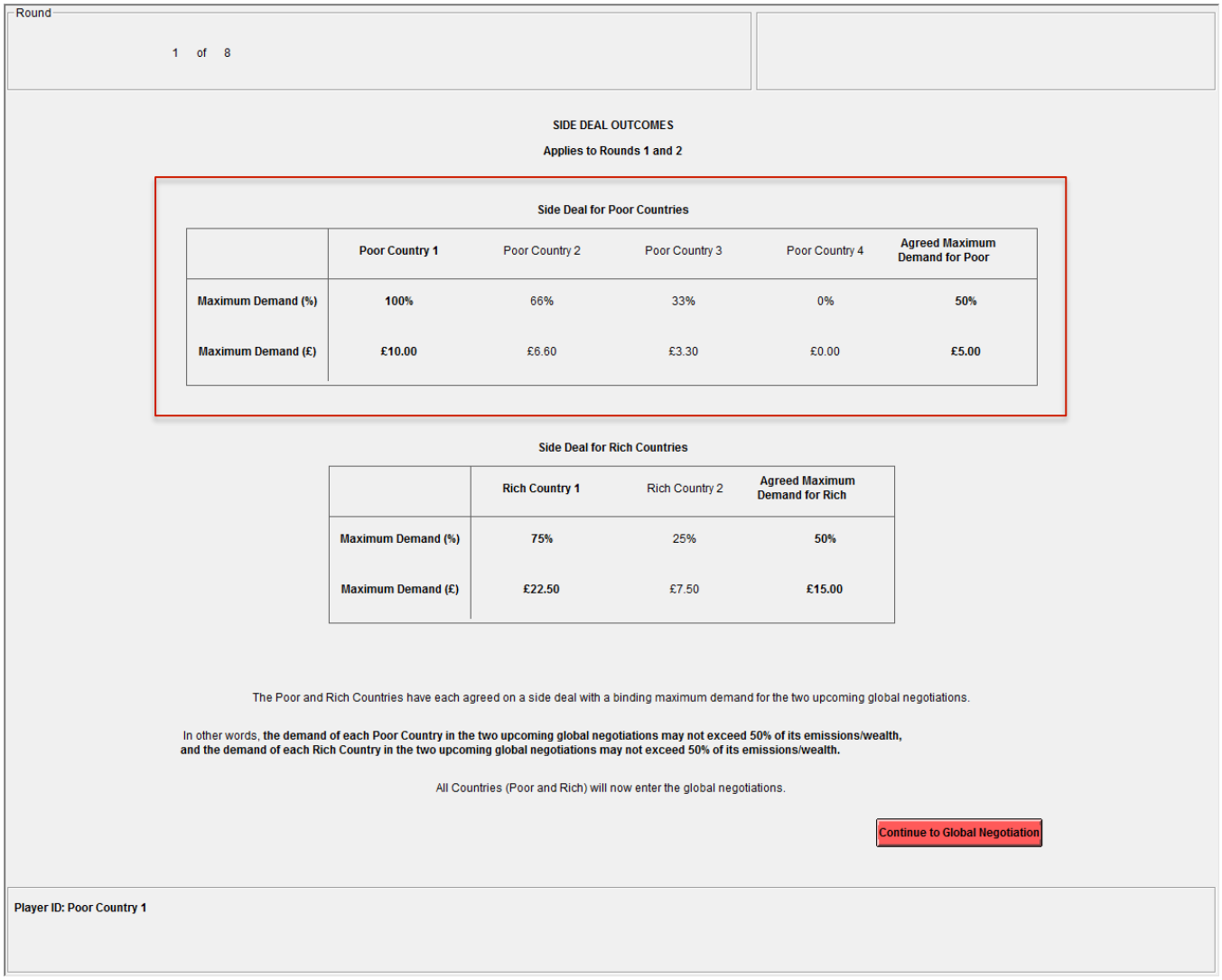
Screenshot 3 I Outcome screen presented to Poor Country 1 in Round 7 of the Global Negotiation if Rich Cou Poor Country 4 demand to keep 32% of their initial wealth/emissions and all other players demand to keep 20 initial wealth/emissions. The Global Demand is less than the Global Target of 30% in Round 7. Each player re demand and negotiations terminate.



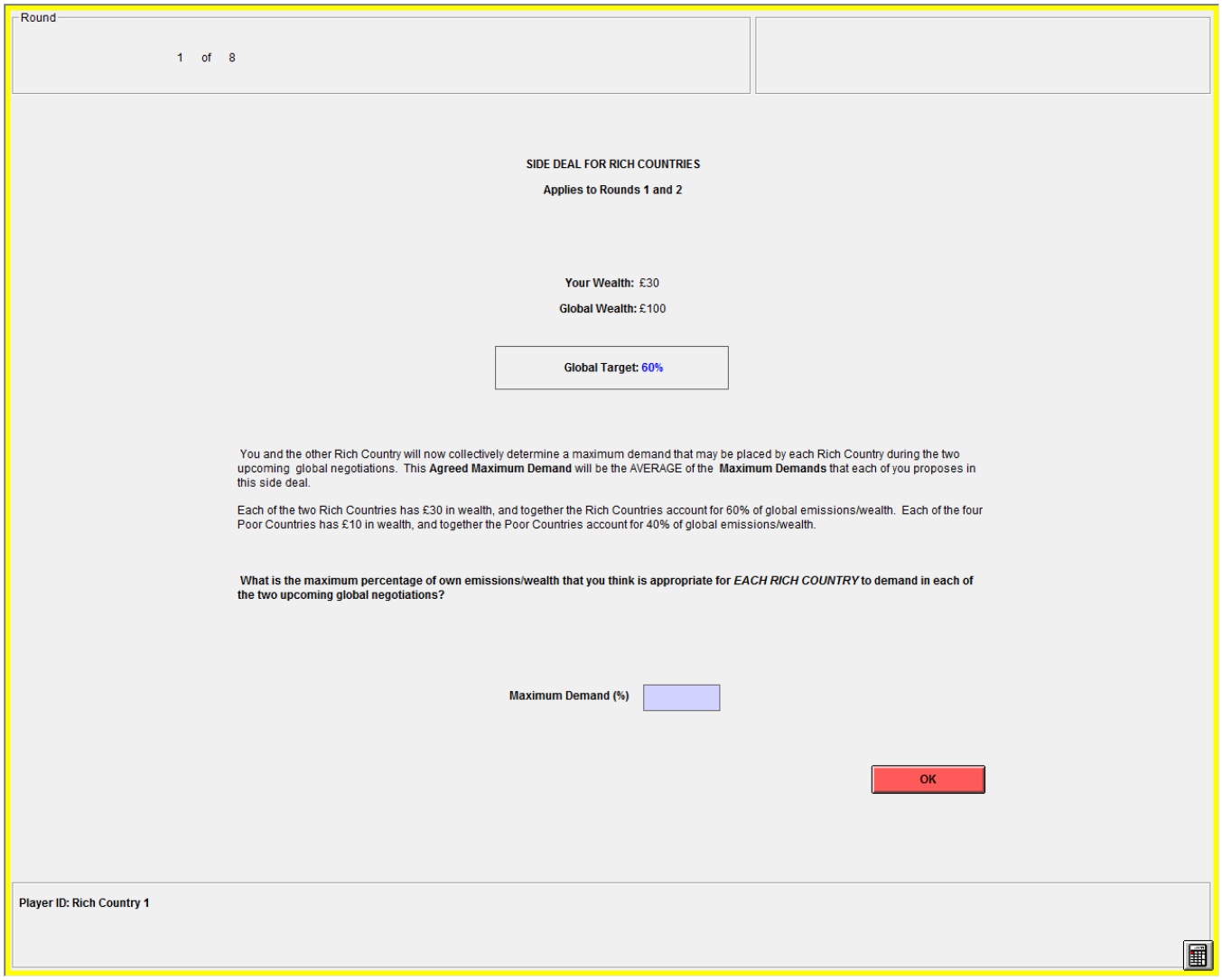
Screenshot 4 I A schematic representation of the stages in treatment ASD. In even-numbered rounds there is stage (Global Negotiation), while in odd-numbered rounds that stage follows a Side Deal stage. The same appli and RSD, except that the Side Deal in those treatments are determined by (and pertain only to) Poor and Rich respectively.



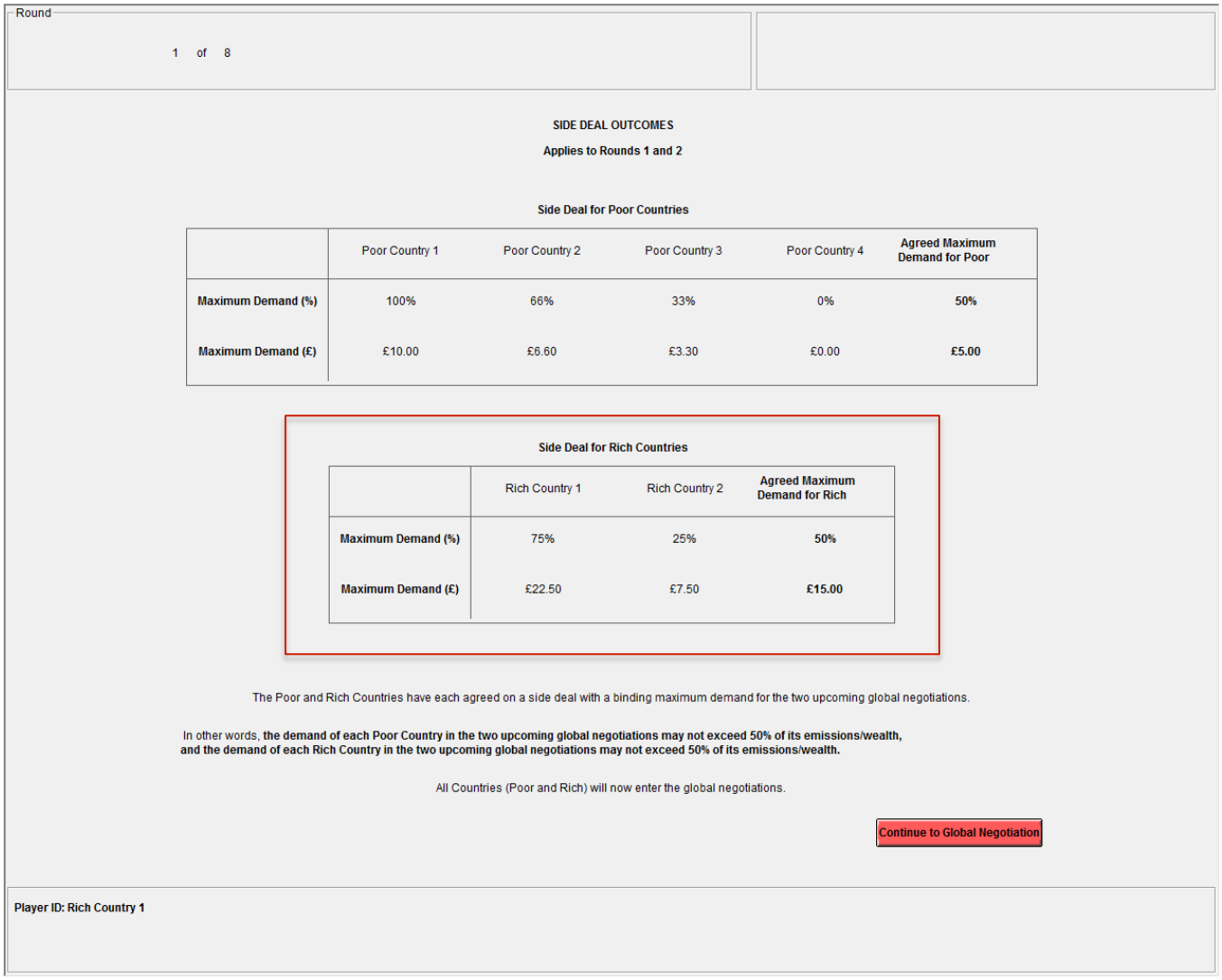
Screenshot 5 I Input screen presented to Poor Country 1 to designate a preferred Maximum Demand in the Poor Side Deal prior to Rounds 1 and 2 of the Global Negotiation.



Screenshot 6 I Outcome screen presented to Poor Country 1 displaying the selected Maximum Demands o players in her group. The red box is included in the Experimental Instructions to highlight the relevant Agreed Demand from the perspective of Rich Country 1, though it does not appear on screen during the experiment Maximum Demands for both Rich Countries and Poor Countries are revealed to all group members prior to the su Global Negotiation stages.

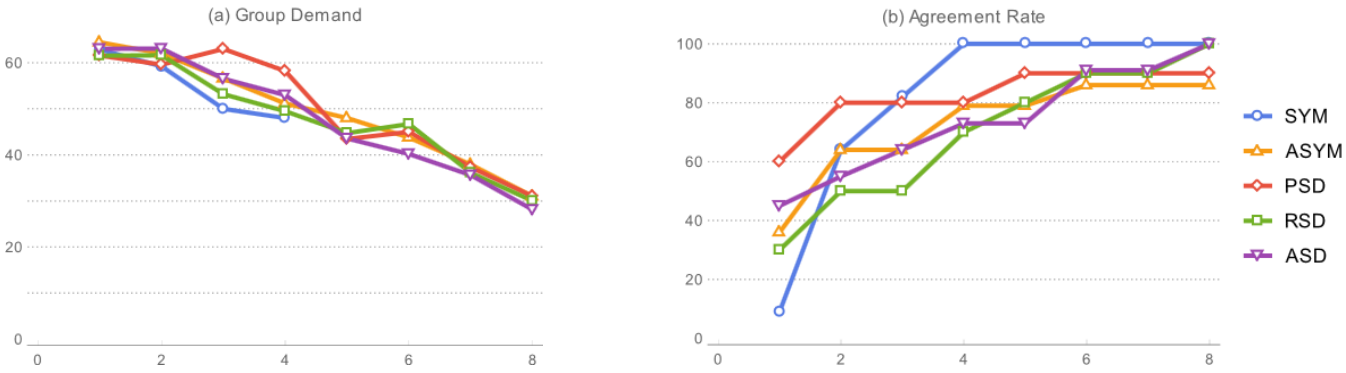


Screenshot 7 I Input screen presented to Rich Country 1 to designate a preferred Maximum Demand in the Rich Side Deal prior to Rounds 1 and 2 of the Global Negotiation.

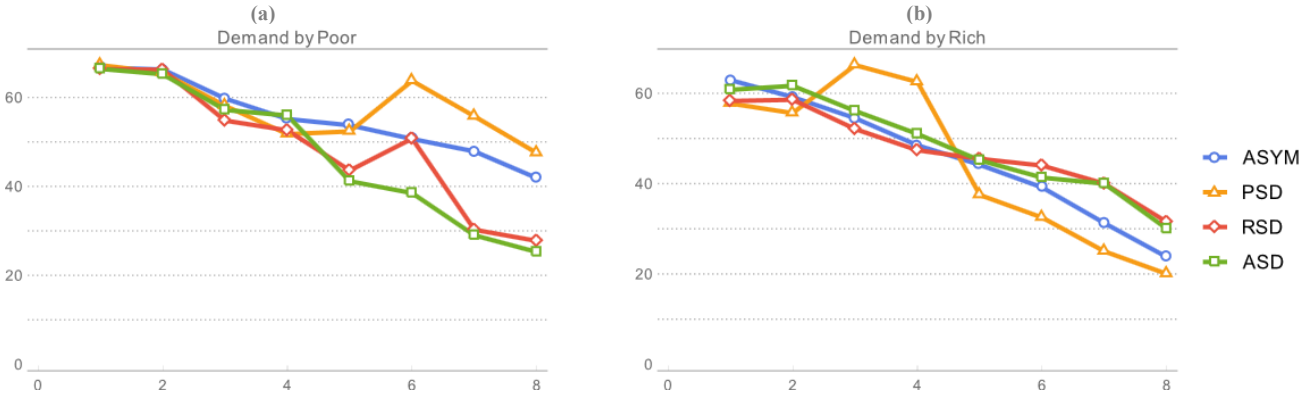


Screenshot 8 I Outcome screen presented to Rich Country 1 displaying the selected Maximum Demands o players in her group. The red box is included in the Experimental Instructions to highlight the relevant Agreed Demand from the perspective of Rich Country 1, though it does not appear on screen during the experiment Maximum Demands for both Rich Countries and Poor Countries are revealed to all group members prior to the su Global Negotiation stages.

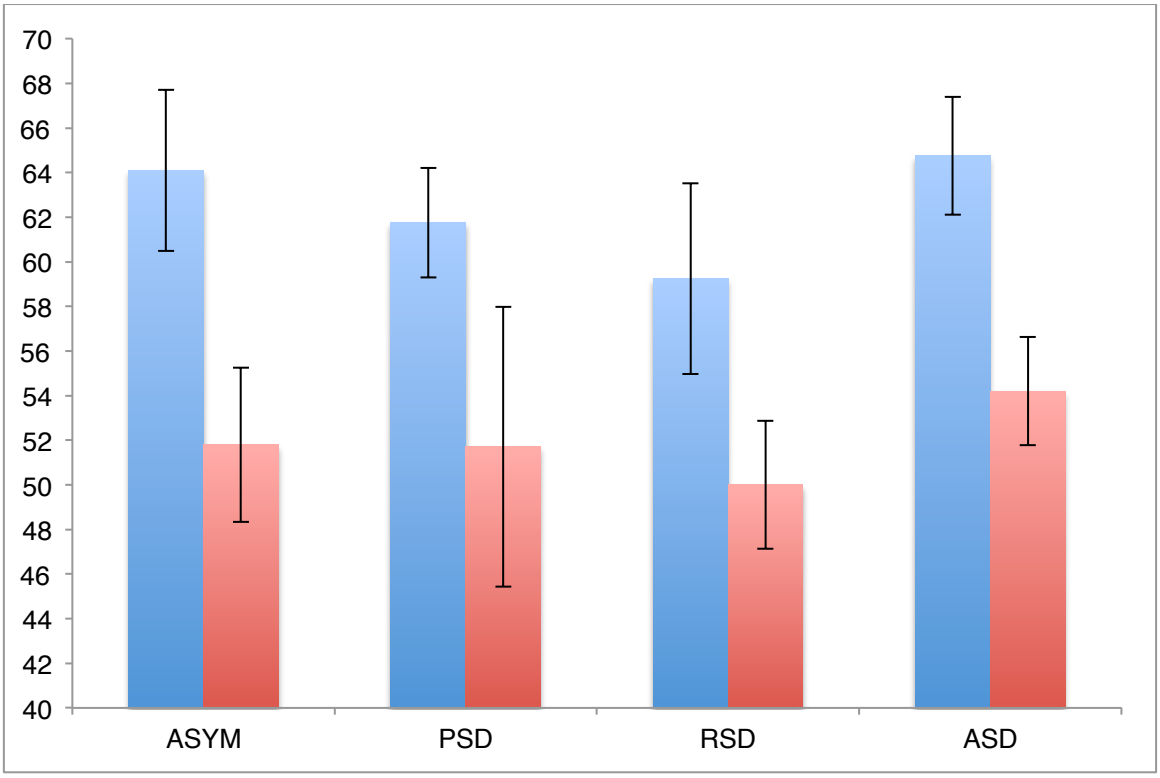
Supplementary Figures



Supplementary Figure 1 | Group demand over time (a) and agreement rate (b). Figure 1a illustrates group demand dynamics, while Figure 1b shows the percent of groups that reached agreement, by round and treatment. Data points in 1(a) should be weighted differently according to the number of groups remaining in the negotiation. For instance, 60% of SYM groups reach agreement in Round 1, so the average group demand for SYM in Round 2 represents the average of the 40% of groups who continued to negotiate in Round 2.



Supplementary Figure 2 | Demands over time by treatment, for the Poor (a) and for the Rich (b). The figure illustrates individual demands over time for both player types.



Supplementary Figure 3 I Average demands (and standard errors bars) by Poor (blue) and Rich (red) negotiators in agreement round of successful groups (i.e. groups who reached agreement in the first two rounds).

Supplementary Tables

Supplementary Table 1 | Agreement velocity (average round in which negotiations terminated) and failures (number of groups that failed to reach an agreement), by treatment

	SYM	ASYM	PSD	RSD	ASD
Velocity	2.455	3.071	2.300	3.400	3.091
	(0.934)	(2.556)	(2.359)	(2.413)	(2.548)
Failures	0	2	1	0	0
Groups	11	14	10	10	11

Supplementary Table 2 | Risk preferences and individual demands

	(1) No Controls	(2) With Controls	(3) With Threshold Control	(4) With Threshold Interaction
Risk	1.680***	1.241**	1.769***	1.541**
	(0.535)	(0.562)	(0.666)	(0.678)
Threshold Round			-3.989***	-5.737***
			(0.550)	(1.015)
Threshold Round * Risk				0.476*
				(0.276)
Constant	52.198***	58.097***	61.459***	62.294***
	(2.270)	(2.631)	(3.105)	(3.069)
Groups	54	54	34	34
Subjects	324	324	204	204
Obs	930	930	810	810

The table displays the results of a panel OLS regression with errors clustered at the group level, where the dependent variable is individual demand. The risk question posed to subjects is based on the incentive-compatible risk preference elicitation gambles in (7, 8). Robust standard errors are reported in parentheses. *Threshold Round* is a dummy equal to 1 if the present round is the threshold round before a decline in the Global Target (i.e. an even round). The number of observations reduces with the threshold control since 20 groups who reach agreement in Round 1 will not experience variation in the *Threshold Round* control and are thus excluded from the regression. Controls include gender, Annex 1 nationality, stated primary motivation, Global Target, and treatment group assignment. ***p<0.001, **p<0.05, *p<0.10

Supplementary Table 3 | Stated risk preferences and individual demands

	(1) No Controls	(2) With Controls	(3) With Threshold Control	(4) With Threshold Interaction
Stated Risk	0.697**	0.728**	0.673*	0.433
	(0.313)	(0.305)	(0.355)	(0.350)
Threshold Round			-3.987***	-6.703***
			(0.550)	(1.505)
TR * Stated Risk				0.496**
				(0.236)
Constant	54.702***	54.510***	64.652***	65.954***
	(1.811)	(2.195)	(2.746)	(2.667)
Groups	54	54	34	34
Subjects	324	324	204	204
Obs	930	930	810	810

The dependent variable in this regression is individual demand. Stated Risk is measured on a scale from 0 to 10 and comes from the general risk question asked in the German Socioeconomic Panel (SOEP; see 12). Robust standard errors are reported in parentheses. Controls include gender, Annex 1 nationality, stated primary motivation, Global Target level, and treatment group assignment. ***p<0.001, **p<0.05, *p<0.10

Supplementary Table 4 | Conditional demands of Poor and Rich (Tobit)

	Poor Demand	Rich Demand
Rich Cooperated	4.074***	0.766
	(1.301)	(2.024)
Poor Cooperated	-0.265	2.420***
	(0.740)	(0.805)
Constant	59.397***	55.995***
	(4.464)	(5.315)
Groups	26	26
Subjects	104	52
Obs	356	178
Controls	Yes	Yes

The table displays the results of a panel Tobit regression, where the dependent variable indicates the percentage demanded o one's initial endowment. The independent variables represent the number of Rich and Poor Country representatives (respectively) who cooperated in the prior round by demanding less than or equal to the Global Target. Controls include gender Annex 1 nationality, stated primary motivation, Global Target, and the difference between the group demand and the target in the prior round of negotiations. There are 26 groups in heterogeneous treatments that negotiated past the first period, and these are the groups considered here. Robust errors are clustered at the group level. Standard errors are reported below estimates in parentheses. ***p<0.001, **p<0.05, *p<0.10

Supplementary Table 5 | Risk preferences and individual demands (Tobit)

	(1) Risk	(2) With Controls	(3) With Threshold Interaction	(4) With Threshold Interaction
Risk	1.659***	1.269**	1.805***	1.575**
	(0.536)	(0.541)	(0.604)	(0.635)
Threshold Round			-4.015***	-5.775***
			(0.563)	(1.613)
Threshold Round * Risk				0.480
				(0.412)
Constant	52.429***	58.142***	61.541***	62.382***
	(2.135)	(2.910)	(3.340)	(3.417)
Groups	54	54	34	34
Subjects	324	324	204	204
Observations	930	930	810	810

The table displays the results of a panel Tobit regression, where the dependent variable is individual demand. The risk question posed to subjects is based on established incentive-compatible risk preference elicitation gambles (7, 8). Robust standard errors are reported in parentheses. *Threshold Round* is a dummy equal to 1 if the present round is the threshold round before a decline in the Global Target (i.e. an even round). The number of observations reduces with the threshold control since 18 groups who reach agreement in Round 1 will not experience variation in the *Threshold Round* control and are thus excluded from the regression. Controls include gender, Annex 1 nationality, stated primary motivation, Global Target level, and treatment group assignment. ***p<0.001, **p<0.05, *p<0.10

CHAPTER VI

CONCLUDING THOUGHTS

Academics, businesspeople, and policymakers alike are increasingly convinced that exogenous variation is key to understanding causal effects of business and government policies and programs. The creation of the Behavioural Insights Team in the UK Government marked the first institutionalization of the experimental methodology within a national government. The United States government has followed suit in an Executive Order by Barack Obama mandating government agencies to seek areas where behavioral and experimental research can improve policymaking, and to do so under the guidance of the new Social and Behavioral Sciences Team.[99] While researchers have recently begun to investigate the prospects for experimental methodology in improving environmental outcomes, the extent of such research in the field of environmental economics is inadequate in relation to its potential scope.

This thesis aims to provide impetus for a movement toward much more prolific implementation of controlled experiments on small and large scales to improve our understanding of human decision making and its consequences for environmental and resource outcomes, as well as the consequences of such outcomes on human welfare. Governments, businesses, and other organizations can benefit from partnerships with academics that seek to gain an increasingly nuanced understanding of the motivations behind the resource-intensive actions or voluntary contribution decisions of customers, employees, and even high-tier decision makers. Experimental results can test and inform theories of behavior, forming a symbiotic relationship that will catalyze knowledge production on important and time-sensitive issues such as climate change, biodiversity loss, and non-renewable resource depletion.

The four experiments presented in this thesis have modestly expanded our understanding of the subtle influences of choice architecture on decision making in a number of contexts. With respect to the behavior of green consumers, reminding individuals of their past environmentally-conscious decisions and encouraging them to be consistent (i.e. to avoid cognitive dissonance) may hold potential for eliciting resource-saving decisions. Such a strategy could be expanded to various contexts, such as food choice (e.g., persuading vegetarians to consume local or seasonal produce), transport (e.g., persuading a hybrid owner to ride her bicycle for short-distance travel or purchase carbon offsets for flights), or

[99] Link to Executive Order: https://www.whitehouse.gov/the-press-office/2015/09/15/executive-order-using-behavioral-science-insights-better-serve-american.

investing (e.g., encouraging environmental donors to invest in green funds or open green bank accounts). While some may argue that the margin for behavioral improvement is diminished for those with strong personal environmental norms, I argue to the contrary that these individuals may be important targets if interventions require that one have sufficient knowledge of or sympathy for environmental causes.[100]

Shifting our attention to employee behavior, provision of salient resource-efficient performance information with targets for improvement proved highly effective in increasing airline captains' fuel efficiency. Moreover, directing charitable incentives at prosocially motivated captains may further enhance efficiency. With new knowledge that risk aversion plays a role in fuel uptake decisions, informational interventions or technologies may be designed to correct inflated subjective probabilities of flight incidents, or to improve certainty surrounding moving variables such as weather and probability of diversions. Again, such knowledge may be applicable to other relevant industries—such as shipping and trucking—and perhaps also to government bodies such as the postal service or the military.

Finally, in the realm of international climate bargaining, side agreements—such as those negotiated between China and the United States as well as China and the European Union prior to COP 21 in Paris—appear to dampen the carbon demands of high-emitting countries when facing a global carbon budget, and to increase the disparity between the concessions that heterogeneous parties are willing to accept. Additionally, the experiment emphasizes the crucial need for high emitters to demonstrate strong commitment and willingness to cooperate in high-N multilateral negotiations with a large degree of heterogeneity, though conditionality stipulations may need to be enacted to minimize free riding by countries with lesser emissions.

Of course, experimentation is no panacea for the behavioral adjustments necessary to combat pressing environmental issues. There are certainly a number of important questions that cannot realistically be resolved (completely) by experiments, and advances in econometric methods have undoubtedly improved the ability of observational data to infer causality and inform solutions. Additionally, qualitative studies continue to shape the way in which we understand the problems at hand. However, most environmental dimensions can benefit in some capacity from experimentation,

[100] Additionally, these individuals may exhibit behaviors consistent with moral licensing, whereby they perform environmental behaviors in one context and therefore do not feel compelled to do so in another (e.g., Tiefenbeck et al., 2013). Policies and programs must take into account potentially harmful spillovers resulting from this phenomenon that may result and take measures to overcome them.

and the more we understand the role of experimentation, the more obvious its potential applications will become.

While the experiments presented here have been implemented solely in developed countries—and despite the stronger presence of experimental methods in development economics—there remains wide scope for experimentation to target decision making in the developing world relating to environmental issues. Such experiments could expand upon our knowledge of (for example) optimal climate adaptation strategies (e.g., Duflo, Robinson, and Kremer, 2011), common pool resource extraction (e.g., Cardenas, Janssen, and Bousquet, 2011), public good provision (e.g., O'Garra, Alfredo, and Schneider, 2015; Jack and Recalde, 2015), policy selection or implementation (e.g., Duflo et al., 2013), improved energy access (e.g., Jack and Smith, 2015), and excess fertility reduction (e.g., Ashraf, Field, and Lee, 2014). Environmental economists may consider teaming up with development and health researchers to bring a stronger and more explicit environmental perspective to such projects.

The field of environmental economics only stands to gain from placing a more pronounced emphasis on experimental research, both to test its own existing theories as well as those from traditional and behavioral economics that speak to environmentally relevant decision making. As noted in *The Stern Review*, while not sufficient to mitigate climate change and resource destruction, behavior change is indisputably an integral component in creating a pathway to an environmentally sustainable future. Thus, research in environmental economics must adopt behavior change as a core tenet of its undertaking to optimize social welfare. Through iterated experimentation across a number of contexts, researchers can begin to paint a clearer picture of human motivation and decision making, creating a more definitive role for choice architecture in improving environmental outcomes.

REFERENCES

Ashraf, Nava, Erica Field, and Jean Lee. 2014. "Household bargaining and excess fertility: An experimental study in Zambia." *American Economic Review* 104 (7): 2210-2237.

Cardenas, Juan-Camilo, Marco Janssen, and Francois Bousquet. 2013. "Dynamics of rules and resources: Three new field experiments on water, forests and fisheries." *Handbook on experimental economics and the environment*, eds. John A. List and Michael K. Price, pp. 319-345.

Duflo, Esther, Jonathan Robinson, and Michael Kremer. 2011. "Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya." *American Economic Review* 101(6): 2350–90.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2013. "Truth-telling by third party auditors and the response of polluting firms: Experimental evidence from India." *Quarterly Journal of Economics* 128 (4): 1499-1545.

Jack, B. Kelsey, and Maria Recalde. 2015. "Local leadership and the voluntary provision of public goods: Field evidence from Bolivia." *Journal of Public Economics* 122: 80-93.

Jack, B. Kelsey and Grant Smith. 2015. "Pay as you go: Pre-paid metering and electricity expenditures in South Africa." *American Economic Review Papers & Proceedings*, 105 (5): 237-241.

O'Garra, Tanya, Katherine Alfredo, and Claudia R. Schneider. 2015. "The influence of social and psychological interventions on collective action for water management: A framed field experiment in India." Working paper.

Tiefenbeck, Verena, Thorsten Staake, Kurt Roth, and Olga Sachs. 2013. "For better or for worse? Empirical evidence of moral licensing in a behavioral energy conservation campaign." *Energy Policy* 57: 160-171.