UNIVERSITY OF CALIFORNIA

Los Angeles

Three Essays on Norms

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Management

by

Jonathan Elliot Bogard

2022

ABSTRACT OF THE DISSERTATION

Three Essays on Norms

by

Jonathan Elliot Bogard

Doctor of Philosophy in Management

University of California, Los Angeles, 2022

Professor Craig R. Fox, Co-Chair

Professor Noah J. Goldstein, Co-Chair

The human desire for belonging and group membership has long been recognized as a fundamental psychological need. From this need comes a tendency for people to look to others' behavior as a clue for how they themselves should act. As a result, learning about the descriptive social norms for behavior can often cause people to assimilate their own behavior in the direction of the social norm. In this research, I explore factors influencing people's perceptions of norms as well as people's reactions to normative information and normative violations. In Chapter 1, we decompose normative comparisons into three separate components, each with their own causal contribution to people's response to such comparisons: Target (the reference group to whom an individual is compared), Distance (how far the individual is from the Target), and Valence (whether the individual has over- or under-performed relative to this benchmark). In establishing these three distinct factors, we are better able to predict how people will respond to receiving normative comparison feedback. In Chapter 2 we use Norm Theory to explain what otherwise appears to be an aversion to utilizing algorithmic recommendations, even when such algorithms obviously outperform human judges. This research endeavor represents but one example of how otherwise puzzling human behavior can be better understood by considering the broader normative context. Finally, in Chapter 3 we document an ironic effect of behavioral interventions on people's perceptions of descriptive social norms. We show that awareness of the presence of a nudge can be enough to lower people's perceived descriptive social norm. Taken together, this body of research seeks to contribute to the theory and understanding of the causes and consequences of perceived descriptive social norms.

The dissertation of Jonathan Elliot Bogard is approved.

Eugene Matthew Caruso

Jana Gallus

Suzanne Bliven Shu

Craig R. Fox, Committee Co-Chair

Noah J. Goldstein, Committee Co-Chair

University of California, Los Angeles

2022

For Mom, Dad,
Vanessa and Sofia.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Lucky graduate students end up in departments where they find an excellent mentor.

Very lucky students end up in departments where they find two or three.

UCLA has given me a lifetime's supply of mentors, role models, intellectual heroes, and friends. Through countless hours of conversation and education, formal and informal, I have become the researcher I am because of their counsel. Their level of sincere, extraordinary care and concern for my well-being—both academic and personal—is at times hard to believe. I tell their stories to anyone who will listen, and I carry them on my shoulders at every moment when developing ideas, conducting research, and crafting papers. Their mentorship and influence will have a permanent mark in my career and life.

Craig Fox, who took a risk on a completely untutored teacher from New Orleans, has one of the fastest, hungriest, and sharpest minds I will ever encounter. He also radiates eagerness and integrity, demonstrating a deep belief in the goodness of others with every interaction he has. Noah Goldstein is a stop-you-in-your-tracks expansive and creative thinker, with an inspiring empirical approach to understanding how people think. He is also among the most generous, benevolent, and conscientious humans I will ever know. Eugene Caruso has an unmatched ability to interrelate ideas in ways that are revelatory and profound, and I spent most of graduate school catching up to quips he casually mentioned off the cuff. He is also as gracious and considerate of others as he is funny. (Very.) Jana Gallus is able to switch from exacting, rigorous attention to detail to sweeping, ambitious theorizing faster and better than anyone I have known. But there are no other scholars of her caliber who are also as warmhearted or thoughtful as she. Suzanne Shu does incisive, world-class theoretical research while simultaneously devoting

herself to applying that research toward making others' lives better—an example I'll spend my career chasing. This ability is matched only by her overwhelming kindness and the fierceness with which she champions others. What's true of all five of them is that they are each unusually talented *and* unusually kind. I'm better for them.

I realized only after completing this dissertation that, by pure happenstance, each of *their* graduate advisors is cited somewhere in this manuscript. How lucky I am to have had the support of giants in the field, themselves standing on the shoulders of giants. The influence of my mentors lives in the words and ideas of this manuscript, and will continue to reverberate through my career. I look forward to reading my future students' dissertations celebrating their work!

I owe my success and my happiness at UCLA to my classmates: Joey and David, academic idols and lifelong friends, as well as the magic, joy, love and electricity of OG Game Night (Ipek, Julia, Pedro, and Sherry) and the BDMers (Malena, Yilin, Ilana, Megan, and Jo).

Most of all, I am grateful to my family, who fill me with love and purpose, and who also often put up with far more than they signed up for. I'm grateful to my parents, my best friends and role models. To my dad, my lifelong partner and my inner voice. And to my mom, my model of how to love others and my moral hero. To Vanessa, who is deeper and more wonderful than anyone deserves, and who grows deeper and more wonderful every day. Her support and patience carried me through graduate school, and her ability to love and challenge me carries us through life. And finally, of course, to Sofia, who has made me happier and more deeply satisfied than I thought was possible, and who brings an intensity, curiosity, sparkle, and kindness to others that I will spend the rest of my life striving toward.

How lucky does one person get to be?

Chapter 1 of this dissertation is a lightly modified version of Bogard, J. E., Delmas, M. A., Goldstein, N. J., & Vezich, I. S. (2020). Target, distance, and valence: Unpacking the effects of normative feedback. *Organizational Behavior and Human Decision Processes*, 161, 61-73. The theorizing, empirical work, and writing was joint with Maggie Delmas, Noah Goldstein, and Stephanie Vezich.

An adapted version of Chapter 2 of this dissertation is currently under academic peer review. The work was devised, executed, and reported in partnership with Suzanne Shu.

Chapter 3 of this dissertation is a Working Paper in which empirical explorations are ongoing. All work to date is shared with Noah Goldstein.

CURRICULUM VITAE

Jonathan E. Bogard

EDUCATION

UCLA Anderson School of Management, Los Angeles, CA *2017-Present*
PhD Candidate in Behavioral Decision Making
Advisors: Craig Fox and Noah Goldstein
Dissertation Committee Members: Eugene Caruso, Jana Gallus, and Suzanne Shu

Brown University, Providence, RI *2005-2009*
B.A. Philosophy, with Honors
Thesis: *Introducing Virtue Ethics to the Problem of Moral Luck*

PEER-REVIEWED JOURNAL ARTICLES

Bogard, J. E., Delmas, M. A., Vezich, S., & Goldstein, N. J. (2020) Target, Distance, and Valence: Unpacking the Effects of Normative Feedback. *Organizational Behavior and Human Decision Processes. 161S.* 61-73.

Krijnen, J., Ülkümen, G., **Bogard, J.E.**, & Fox, C.R. (2021) Lay Theories of Financial Well-being Predict Political and Policy Message Preferences. *Journal of Personality and Social Psychology.*

Milkman et al. [incl. **Bogard, J. E.**] (2021) A Mega-Study of Text-Based Nudges Encouraging Patients to Get Vaccinated at an Upcoming Doctor's Appointment. *Proceedings of the National Academy of Sciences.*

Lerner et al. [incl. **Bogard, J. E.**] (2021). Patient Portal Reminders for Pediatric Influenza Vaccinations: A Randomized Clinical Trial. *Pediatrics*.

Szilagyi et al. [incl. **Bogard, J. E.**] (2021) Effect of Personalized Messages Sent by a Health System's Patient Portal on Influenza Vaccination Rates: A Randomized Clinical Trial. *Journal of General Internal Medicine.*

Milkman et al. [incl. **Bogard, J. E.**] (2022) A 680,000-Person Megastudy of Nudges to Encourage Vaccination in Pharmacies. *Proceedings of the National Academy of Sciences.*

CONFERENCE PRESENATIONS

Heuristics & Biases in Judgments of Economic Inequality, with Craig Fox and Colin West

- Society for Judgment and Decision Making Conference, New Orleans, November 2018
- Society for Consumer Psychology, Huntington Beach, March 2020
- Association of Psychological Science Convention, Chicago, May 2020
- Society for Consumer Psychology, Hawaii, July 2021

Averse to Algorithms or Averse to Uncommon Decision Procedures? with Suzanne Shu

- Association for Consumer Research, Virtual, October 2020
- Society for Judgment and Decision Making, Seattle, November 2021

Enhancing Probability Sensitivity through Outcome Simulations, with Craig Fox

- Behavioral Science and Health Symposium, Virtual, December 2020
- Society for Judgment and Decision Making Conference, Virtual, December 2020
- Association for Psychological Science Convention, Virtual, June 2021
- Association for Consumer Research, Seattle, October 2021
- Symposium lead and paper presentation at the Society for Personality and Social Psychology, Virtual, February 2022

Perceptions of Inequality: The Moral and Political Causes and Consequences, with Dylan Wiwad, Christopher To, and Craig Fox

- Symposium lead and paper presentation at the Society for Personality and Social Psychology, Virtual, February 2021

Lay Beliefs about Changes in Financial Well-being Predict Political and Policy Message Preferences, with Job Krijnen, Gülden Ülkümen, and Craig Fox

- Society for Consumer Psychology, Hawaii, July 2021, *postponed*

The Implied Exclusivity Effect: Promoting Choice Through Reserved Labeling, with Craig Fox and Noah Goldstein

- Association for Consumer Research, Seattle, October 2021
- Society for Judgment and Decision Making, Seattle, November 2021

Intrinsic Motivation, Pay-for-performance, and Organizational Missions

- Psychology of Technology Conference, November 2021

The Stanford University Summer Institute in Political Psychology, 2019

CONFERENCE POSTERS

Thou Shalt Bear True Witness: The Moral and Behavioral Consequences of Encouraging Honesty versus Discouraging Dishonesty, with Eugene Caruso, Alex Shaw, Shoham Choshen-Hillel, and Ashley Feinsinger

- Life Improvement Sciences Conference, Virtual, 2021

Enhancing Probability Sensitivity through Experiential Simulations of Outcomes, with Craig Fox

- The European Association for Decision Making, SPUDM, Virtual, August 2021

PROFESSIONAL EXPERIENCE

Collegiate Academies, *New Orleans, LA* *2011-*
Mathematics teacher, Dean, Teacher trainer, Director of Teacher Coaching *2017*

L.W. Higgins High School, *Marrero, LA* – Teach For America | New Orleans *2009-*
Teacher, Geometry *2011*

INTRODUCTION

The Need to Belong and the Power of Norms

In 2016, an international team of developmental psychologists brought five-year-old children into their lab for an experiment (Misch et al., 2016). The researchers would sit each child in a circle with four puppets (controlled by the researchers), then start a conversation between the child and the puppets in order to relax the young participants. After a few minutes of conversation, the researchers would open a box containing several scarves and give two of the puppets a green scarf to wear and the other two puppets a yellow scarf. The researchers then gave the child either a green or yellow scarf, thereby creating a pair of in-group puppets and a pair of out-group puppets with the child. After a decoy task and a cover story, each child would overhear a secret that was exchanged between either the pair of in-group or out-group puppets. Then a fifth puppet would enter the scene and bribe the child to reveal the secret in exchange for a sparkly sticker. The researchers were interested in whether or not the child would divulge the secret in exchange for a sticker, depending on whether the secret was told between the in-group or out-group puppets. In other words, the researchers wanted to know if children would be more willing to make a personal sacrifice to protect a member of their in-group. They found that the children were over 50% more likely to keep the secret, at the personal cost of a coveted sticker,[1] if the puppets were part of the child's in-group rather than out-group.

So great and so innate is the human desire to belong to and fit in with a group. In fact, psychologists have identified the desire to belong as a fundamental psychological need

[1] The author, from recent field observation, can attest to the extraordinary intensity with which children value stickers!

(Baumeister & Leary, 1995; Maslow, 1943; Schachter, 1959). It is from this need, perhaps, that people derive a tendency to use others' actions to guide their own decisions. Applying a "popular = good" heuristic (Cialdini et al., 1991), people often rely on the behavior of others as a heuristic for figuring out what is best for themselves (Cialdini & Goldstein, 2004). As Cialdini puts it, "We view a behavior as more correct in a given situation to the degree that we see others performing it" (Cialdini, 2021). Decades of research have documented that, as a consequence of this tendency, providing people with normative information can lead them to assimilate into the common behavior, from drinking alcohol (Lewis & Neighbors, 2006) to conserving water (Brent et al., 2015), from donating to charity (Shang & Croson, 2009) to paying taxes (Behavioural Insights Team, 2012). So called "norms nudges"[2] owe much of their potency, perhaps, to this fundamental need to fit in and belong.

This dissertation contains three essays, which all fundamentally consider the components and consequences of people's perceptions of *descriptive social norms*. Descriptive social norms are the behaviors and opinions common for a group of people. They are the "rules and standards that are understood by members of a group, and that guide and/or constrain social behavior without the force of laws"(Cialdini & Trost, 1998). People seem to have a nearly automatic ability to apprehend social norms by briefly observing group behavior (Nolan et al., 2008), but they also learn about norms from summary information and institutional signals (Tankard & Paluck, 2016). Throughout, I will distinguish between behavior that is common (the descriptive social norm) and behavior that individuals *believe* is common (the *perceived* descriptive social norm). Often, but not always, these two concepts are overlapping.

[2] This term originates, to my knowledge, with Bicchieri, C., & Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public choice*, 1-22.

Chapter 1, published in *Organizational Behavior and Human Decision Processes*, begins from the observation that, while many studies have documented the expected effects of norms nudges, many others find null or even negative effects of norms nudges. Contemporary theorizing about social proof cannot accommodate these conflicting results. In trying to explain the variable response to normative information, we identify the independent causal effect of three distinct components of normative comparisons: the target (to whom a person is compared), the distance (how far a person is from that target), and the valence (whether a person has under- or over-performed relative to that benchmark). In so doing, we hope to add greater nuance to discussions of social proof, construing it not as a monolith but instead as a rich, multi-faceted packet of information leading to complex behavioral responses.

In Chapter 2, currently under peer review, we take on a question that has puzzled behavioral scientists since at least the 1980s: Why do humans often ignore the judgment of demonstrably better algorithms in favor of human judgments (Arkes et al., 1986)? Igniting a firestorm of recent research in a similar vein, Dietvorst and colleagues suggest that humans have a fundamental aversion to using algorithms (Dietvorst et al., 2015). At odds with this impressive preponderance of scientific evidence, however, is the obvious extent to which algorithms have been fully integrated into daily living, influencing everything from the news articles we read each morning, to the traffic lights we encounter on our way home from work. Nevertheless, it is not hard to get an intuitive feeling for what Dietvorst was onto—reflect for a moment about whether you would rather have a life-or-death medical decision made by the world's best physician or the world's best algorithm. So what explains both an apparent aversion to algorithms as well as a proliferation of algorithms in our daily lives? We use Kahneman and Miller's (1986) Norm Theory to argue that what often seems to be an aversion to algorithms is actually an aversion to *algorithms that are unconventional* (i.e., against the norm). It is the fact that they are unconventional, we show, that creates most of the apparent aversion to algorithms. Once an algorithm becomes the norm, much of the aversion melts away (and even reverses). This endeavor represents just one way in which otherwise puzzling human behavior may be better understood by considering the broader normative context.

Finally, in Chapter 3, we document a novel factor influencing people's perceptions of the descriptive social norm: the presence of a behavioral intervention (i.e., a nudge). We show that people engage in social sensemaking, inferring that a nudge was implemented by a choice architect in response to a specific problem. This inference in turn leads people to presume that there must be a descriptive social norm opposed to the desired behavior. In fact, this inference is so strong that it even affects norms nudges themselves. That is, even nudges that operate by providing favorable information about the descriptive social norm are subject to this backfiring effect. This creates a "negative social proof" element of any nudge for which this is true. We speculate that this may help explain why nudges with a strong scientific evidence base sometimes fail to have the expected results when implemented in the field.

The cross-cutting theme of this dissertation is the power of norms, both in shaping and explaining human behavior. Altogether, the goal of the present research is to advance our theory and understanding of the causes (Chapter 3) and consequences (Chapter 1 and Chapter 2) of perceived descriptive social norms. I turn now to Chapter 1, decomposing normative comparisons.

CHAPTER 1

TARGET, DISTANCE, and VALENCE: UNPACKING THE EFFECTS OF NORMATIVE FEEDBACK

Jonathan E. Bogard [a]

Magali A. Delmas [a,b]

Noah J. Goldstein [a]

I. Stephanie Vezich [a]

[a] Anderson School of Management, University of California, Los Angeles
[b] Institute of the Environment and Sustainability, University of California, Los Angeles

We would like to thank the California Air Resources Board (contract # 10-332), the National Science Foundation (grant # 1257189), the Morrison Family Center for Marketing Studies and Data Analytics, and the UCLA Anderson Easton Technology Management Center for helping fund this research.

TARGET, DISTANCE, and VALENCE:
UNPACKING THE EFFECTS OF NORMATIVE FEEDBACK

ABSTRACT—People constantly receive information about their performance relative to others. Estimating these effects is complicated because, as we show, normative feedback includes several dimensions: Target (e.g., a reference group of average versus exemplary performers), Distance (e.g., being near versus far from a benchmark), and Valence (e.g., being better or worse than the benchmark). In Study 1, we randomly assign households to receive no feedback or feedback comparing their energy consumption to either their average or most efficient neighbors. Households compared to average neighbors decreased electricity usage by 6%, but those compared to efficient neighbors *increased* consumption by 4%. We decompose these effects into the separate influences of Target, Distance, and Valence. In Studies 2 and 3a-c, we randomly assign normative feedback to isolate the independent effects of Distance and Valence. Additionally, we find evidence for the mediating effect of motivation: The more dispiriting the feedback, the worse the subsequent performance.

Keywords: normative influences; social comparison feedback; residential energy use; energy conservation; healthy behaviors.

1. Introduction

Suppose you are a manager wanting to motivate your employees to work harder. You have heard from your colleagues at other firms that they use a quarterly leaderboard to motivate their staff through peer comparisons, so you decide to do the same. Along with providing individual performance feedback in a report to each employee, the report also compares the employee's performance relative to their peers. But which peers should you choose for this report? For example, should you compare each employee to the median team performance? Or should you compare each employee to a group of high performers from the team, such as the top 20%? If you provided the median comparison, someone in the 52^{nd} percentile would be two points ahead of the reference group. But if you provided the top-performer comparison, that very same person would be 28 percentile points *behind* the reference group. How will your selection influence how employees regard the feedback, and how will it influence their motivation and subsequent performance? Which target is going to be most effective at increasing performance next quarter?

In this paper, we take up the question of how normative feedback influences subsequent motivation and performance. We study this across multiple domains. Our first experiment was a field study that utilized appliance-level household energy consumption feedback wherein participants were randomly assigned to receive no normative feedback or true normative feedback comparing them to either their average or their most efficient neighbors. The second field experiment provided false feedback to participants who had downloaded a pedometer application on their cell phones, randomly assigning participants to receive feedback that they were either very close to or very far behind the top performers in their demographic cohort. In the final set of three experiments, using the same pedometer domain as well as a word-search task and a typing test, we randomly assigned participants to receive feedback that they had performed either better or worse than a reference group, and measured subsequent motivation and performance. The purpose of the current investigation is to demonstrate the independent effects of the previously-hard-to-disentangle distinct elements of normative feedback: the choice of reference group (e.g., average versus exemplary performers; “Target”), the impact of being near versus far from a reference group (“Distance”), and being better or worse than the reference group (“Valence”).

2. Social Norms Feedback

People are generally poor at estimating their relative standing in a variety of performance domains (e.g., Burson, 2007). Perhaps for this reason, social norms feedback—information about one’s attitudes or behaviors relative to those of a relevant social group—has gained particular interest (e.g., Bollinger & Gillingham, 2012; Cialdini & Goldstein, 2004; Cialdini & Trost, 1998; Schultz, 1999; Schultz et al., 2007). Normative information has been shown to powerfully shape behavior (Berkowitz, 1972; Goldstein et al., 2008; Sherif, 1936), even when people believe these comparisons have little impact on their personal choices (Cialdini et al., 1991; Nolan et al., 2008). Social comparison theory suggests that norms are powerful because comparing oneself to the norm establishes the appropriate level of that behavior (Festinger, 1954).

Large-scale field studies of thousands of households demonstrate the effectiveness of normative feedback in reducing household energy consumption (Allcott, 2011; Allcott & Rogers, 2014). In these studies, residents receive feedback about their consumption relative to both the average and the most efficient 20% of consumers. This combination provides residents both a benchmark of how their neighbors are doing overall with the average reference group, and a more high-performing goal to strive toward with the efficient reference group. Their findings show that this combined approach is effective, but they do not allow us to isolate the individual impact of each reference group. That is, we do not know how results might differ if residents only see feedback relative to their average neighbors or their highest-performing neighbors. Further, previous studies are limited in testing the independent ways that the valence of the feedback and the distance from the reference group separately impact behavior. Thus, we set out to investigate the independent effects of Target, Distance, and Valence in contributing to the overall influence of normative feedback.

2.1. Target

Based on past findings, highlighting a high-performing reference group (e.g., efficient neighbors) could have clear advantages. Much of the goal-setting literature suggests that setting harder to reach ("stretch") goals results in greater and more sustained behavior change toward that goal (Kerr & Landauer, 2004; Locke & Latham, 2006). Notably, stretch goals have been recommended specifically in the domain of sustainability (Manning et al., 2006), although they have not yet been tested empirically. If being compared to the most efficient neighbors is similarly aspirational, we would expect that a higher-performing Target should be more effective. However, the opposite relationship between Target and subsequent performance is also plausible. For instance, response to normative feedback is shown to depend on people's attitudes toward (Göckeritz et al., 2010; Neighbors et al., 2010) and beliefs about (Jachimowicz et al., 2018) the reference group. If people do not admire or identify with the higher-performing reference group, or if they believe the high-performing group is exceptional and therefore less relevant to themselves (Alicke et al., 1997), a social comparison against a high-performing reference group may have little influence on them, and thus they may not increase subsequent effort. On the contrary, if people are motivated to match (or even exceed) the average group, we would expect those compared to an average reference group to improve subsequent performance more than those compared to top performers. Thus, we might expect that the effect of Target group will depend critically on people's attitudes toward the specific high-performing group to which they are being compared. For these reasons, we expected the selection of a Target group to exert an independent influence on performance but, depending on individual differences, it was less clear which of the following two competing hypotheses was more likely:

H_{1a}: People who receive normative feedback compared to average-performing peers will improve subsequent performance *more* than those compared to high-performing peers.

H_{1b}: People who receive normative feedback compared to average-performing peers will improve subsequent performance *less* than those compared to high-performing peers.

We test these hypotheses specifically in our study of household energy consumption in Study1.

2.2. Distance

In a natural experiment involving students in a Massive Open Online Class, Rogers and Feller (2016) find that the further behind a person's essay score is from the score of a high-performing peer whose essay they graded, the lower that person's course performance. This suggests that being further behind a benchmark will result in worse subsequent performance. Other work has shown that comparisons to reference groups that are seen as unattainable are demoralizing and thwart effort (Lockwood & Kunda, 1997). Lockwood and Kunda elsewhere found that superior role models can be demotivating when individuals are reminded that even their personal best falls short of the role model's achievement; in such cases there is little incentive to exert effort toward those achievements (Lockwood & Kunda, 1999). Additionally, research has demonstrated that motivation increases as one advances close to goal attainment (Heath et al., 1999; Kivetz et al., 2006; Liberman & Förster, 2008). Thus, a reference group that lies closer to one's current performance may be more motivating than a reference group that is farther away. Given this, in the present experiment we hypothesize that:

> **H_2: People who receive normative feedback that they are closer behind a high-performing reference group will improve subsequent performance more than those who receive normative feedback that they are further behind a high-performing reference group.**

We test this hypothesis both in the context of true feedback of household energy consumption (Study 1) and with false feedback regarding the number of steps taken in a week (Study 2).

2.3. Valence

It is an open question whether being told that one overperformed or underperformed relative to a reference group would lead to greater subsequent performance. On the one hand, there is reason to think that positively valenced normative feedback will lead to considerable performance improvement. For instance, labelling people as top performers relative to peers has been demonstrated to increase doctors' adherence to medical guidelines (Meeker et al., 2016) and the chances that someone votes in an upcoming election (Tybout & Yalch, 1980). In the current research, we pair descriptive feedback with an injunctive norm, which has been shown to also improve performance among those who perform comparatively well (Schultz et al., 2007). This leads us to the following hypothesis:

> **H_3: People who receive positively valenced normative feedback will improve subsequent performance more than those who receive no normative feedback.**

It is noteworthy that Schultz et al. (2007) found an overall backfiring effect of positively valenced normative feedback in the absence of an injunctive norm, but no such backfiring when an injunctive norm was included. In each study of the current investigation, there is either an explicit (Study 1) or more implicit (Study 2 and Study 3a-c) injunctive norm present to mitigate such a backfiring (Asensio & Delmas, 2015). Coupling injunctive and descriptive norms messaging has been shown (e.g., Cialdini et al., 1991b) to be more effective than either in isolation.

The outcome of negative feedback for underperformers may be harder to predict. For instance, Schultz et al. (2007) found that underperformers receiving negatively valenced feedback had an overall improvement in their energy conservation. However, there is other research suggesting that the opposite pattern may emerge: Receiving negatively valenced peer comparison feedback can be demoralizing and hence lead to diminished performance. For example, the Self-Evaluation Maintenance model posits that when people find out that others are performing better than they are on a given task, they start to view the task as less important to their self-definition, which in turn causes them to exert less effort in that domain (Tesser, 1988, 1991; Tesser & Campbell, 1983). Moreover, just as internalizing a positive identity as an overperformer can lead to increased effort in that domain (as in Meeker et al., 2016), internalizing an identity as an underperformer could similarly lead to lower subsequent effort and thus worse performance. Given these conflicted findings, it is an open question how those who receive negatively valenced normative feedback will perform in response to that feedback, setting up two competing hypotheses:

H_{4a}: People who receive negatively valenced normative feedback will improve subsequent performance more than those who receive no normative feedback.

H_{4b}: People who receive negatively valenced normative feedback will improve subsequent performance less than those who receive no normative feedback.

We test the effects of Valence both in the context of household energy consumption (Study 1) and in several other domains (Studies 3a-c).

3. Contribution of the Present Investigation

Using normative feedback that is true (Study 1) and false (Study 2 and Studies 3a-c), we begin to establish three distinct factors often confounded when considering the effects of social comparison feedback: Target reference group (average versus top performers), Distance (how far one is from the Target), and Valence (whether one has over- or under-performed relative to the Target). In Study 1, we provide true feedback compared against randomly assigned Targets in a field study of household energy consumption. Additional analyses show associations with Distance and Valence of feedback. Then in Study 2 and Studies 3a-c, respectively, we randomly assign Distance and Valence feedback to establish their independent effects. We consider the effects of normative feedback in light of explicit (Study 1) and implicit (Study 2 and Studies 3a-c) injunctive norm messages. Altogether, the purpose of the current investigation is to establish the existence of three separate dimensions of normative feedback. In drawing out the independent effects of each of these factors, we move toward helping managers, organizations, and policymakers optimize their use of normative feedback by understanding how the impact of these comparisons varies depending on the recipients' placement within the distribution as well as their individual characteristics.

4. Study 1: Main Effects of Comparison Target on Household Energy Consumption

Recent studies suggest that providing people with better information about their energy consumption can induce energy conservation (Delmas et al., 2013; Delmas & Lessem, 2014; Gillingham & Palmery, 2014; Karlin et al., 2015). In Study 1, we randomly assigned households to receive normative feedback relative to either their average or their most efficient neighbors. This allows us to measure how the choice of Target reference group for social comparison affects subsequent performance. However, in selecting different Targets, we also created natural variation in both the Valence of the feedback and the Distance from the benchmark. For instance, two households both compared to their average neighbors were technically in the same experimental condition but, depending on performance, one might have been slightly below the average whereas the other might have been considerably above the average. Put differently, consider two households, both in the 75th percentile of all homes, that were assigned to different experimental arms in our study. Beyond being compared to different Targets, they also received very different Valence and Distance feedback, one household being considerably ahead of the (average) benchmark and the other being slightly behind the (top-performing) benchmark. This natural variation enabled us to measure the associations with Valence and Distance in randomizing participants to different Target comparisons.

4.1. Methods

We built an intelligent wireless sensor network to provide households with real-time access to detailed, appliance-level information about their home electricity consumption (Chen et al., 2015). We experimentally manipulated normative messages that different households saw, comparing them to different reference groups (average vs. efficient) or a no-feedback control.

Our sample consisted of 101 households in a 1,102-household graduate-student family housing community in Los Angeles, California. The occupancy of each household ranged from 1 to 6, with the number of children per household ranging from 0 to 4. Household size ranged from 1 to 3 bedrooms. Descriptive statistics are listed in Table 1-1.

Table 1-1: Descriptive statistics

	Experiment Participants				**Population**		
	Mean	**Std. Dev.**	**Min**	**Max**	**Mean**	**Std. Dev.**	**p-diff**
Hourly kWh	8.75	5.55	3.66	47.23	8.28	3.62	0.24
Square feet	859.24	109.88	595	1035	868.72	98.40	0.36
Floor	2.14	0.78	1	3	2.08	0.79	0.46
# children	0.57	0.89	0	4	0.51	0.76	0.43
Environmental Organization	0.07	0.26	0	1	NA	NA	NA
Observations	101 households				1102 households		

Participating households did not differ significantly from other households in the housing community in terms of average daily electricity consumption, square footage, what floor they lived on, or number of children (all *p*s greater than .2) as described in Table 1-1. The housing characteristics of our sample were also generally typical of the broader population. For a more thorough treatment of the generalizability of our sample, see the Appendix note N1. Despite the similarities on observed characteristics, we cannot rule out that there are differences on unobserved dimensions with those who did not volunteer to participate, raising the possibility of selection-into-sample bias.

4.2. Procedure and Timeline

Average daily temperature over the course of the field experiment, which ran from September to April, was 60.5 degrees Fahrenheit (*SD* = 6.8). For approximately two months prior to the start of the experiment, as a baseline, we observed households' 15-minute kilowatt-hour (kWh) electricity consumption but did not provide energy use information or any messaging. Households were then randomly assigned to a condition of (a) no-treatment control, (b) Average Reference Group treatment, or (c) Efficient Reference Group treatment.

The treatment started on September 30 and lasted until April 16. During the treatment period, the control households continued to have their consumption tracked every 15 minutes but did not receive any messaging, as before. In contrast, the Average Reference Group treatment and Efficient Reference Group treatment households continued to have consumption tracked every 15 minutes but also gained access to a personal web dashboard and received weekly email reminders to visit the dashboard. In both treatment conditions, residents received feedback on their electricity consumption by month, day, and hour. Feedback could be shown on the dashboard at both the aggregate and the appliance level. Residents could view a pie chart demonstrating the proportion of total energy used by each appliance category (heating and cooling, lighting, plug load, dishwasher, refrigerator, and other kitchen electricity use; see Appendix Figure A1-1 for a screenshot of the weekly email and the website dashboard). This appliance-level information, which was available to both treatment groups but not the control, may represent a considerable improvement upon an aggregate monthly bill; such highly granular feedback may enhance the efficacy of normative feedback, offering clear priorities to those who are motivated to improve their conservation.

Finally, residents in both treatment conditions received feedback about their energy consumption over the past week relative to their neighbors; this is where the manipulation took place. This feedback was sent by email and was also available on each participant's personal dashboard. In both treatments, residents were told that they consumed a certain percentage more than or less than the neighbors in their reference group. This feedback was accompanied by an environmental and children's health message: Residents were told how many pounds of pollutants they contributed to or avoided over the past week (depending on Valence of feedback) relative to the neighbors in their reference group, citing the relationship of these pollutants with health impacts such as childhood asthma and cancer. This injunctive message was selected based on pre-testing that showed environmental and childhood health messaging, especially among parents, to be the most compelling reasons for conservation. Equivalent pounds of air pollutant emissions were calculated using emission factors from the Emissions and Generation Resource Integrated Database (eGRID), maintained by the United States Environmental Protection Agency, based on the Los Angeles Department of Water and Power electricity fuel mix.

In the Average Reference Group treatment, personal feedback was provided relative to the mean total consumption across all neighbors in the sample during the week prior. In the Efficient Reference Group treatment, personal feedback was provided relative to the 20th percentile of total consumption (i.e., the 80th percentile of conservation) across all neighbors in the sample during the week prior. In pre-testing, a focus group of participants indicated that comparisons to the 90th percentile felt unattainably efficient. We instead chose the 80th percentile to match the cutoff used by OPower, a utility company that reports customers' energy consumption using social comparisons to the 80th percentile. Each week, residents could check their performance compared to the reference group to which they were randomly assigned. Participants in the Efficient [Average] Reference Group condition read:

> "Last week you used X% more/less electricity than your efficient [average] neighbors. Over one year, you are adding/avoiding Y pounds of air pollutants which contribute to health impacts such as childhood asthma and cancer."

In addition, definitions of the respective reference groups were provided below each of the messages. An efficient neighbor was defined as the "20% most energy efficient neighbors in similar-sized apartments," whereas an average neighbor was defined as the "average usage in similar-sized apartments." Participants in the control condition received no feedback messages at all.

4.3. Results: Main Effect of Target

We were first interested in investigating the effect of our two messaging treatments on 15-minute-interval household energy consumption, in $\tau=E[Y_{it}(1)-Y_{it}(0)]$, where $Y_{it}(1)$ and $Y_{it}(0)$ represent household i's electricity use at time t if the households were treated and were not treated, respectively (Rubin, 1974). We chose to use 15-minute consumption because it allows us to control for important differences in time-of-day use. We employed a difference-in-difference estimator, which models energy use conditional on post-treatment dummy (*P)*, a messaging treatment dummy (T_{eff} and T_{avg} for those in the Efficient and Average Reference Group, respectively), and their interaction ($P*T_{\mathrm{avg}}$, $P*T_{\mathrm{eff}}$). Hence, we estimate the following model:

$$y = \beta_0 + \beta_1 P + \beta_2 T_{avg} + \beta_3 T_{eff} + \beta_4 (P * T_{avg}) + \beta_5 (P * T_{eff}) + \upsilon + \varepsilon$$

We were chiefly interested in the interaction terms, β_4 and β_5, estimating the effects of each treatment on energy consumption compared to the no-feedback control. Note that we also incorporated a standard set of controls (υ) to account for cyclical time and weather factors, along with demographic factors, gathered through a survey before the start of the experiment, that can greatly impact energy consumption (Grønhøj & Thøgersen, 2011). These factors included: number of children, apartment square footage, floor within building, membership in an environmental organization, hourly temperature, presence of daylight savings time, week in the study, day of the week, and hour of the day. With this estimation strategy, we compare the effect of messaging on change from baseline for the treatment (Average Reference Group or Efficient Reference Group) versus control groups.

Because we had substantially more time periods than individual households, rather than using an ordinary least squares estimator with standard errors clustered at the household level, we used the more efficient feasible generalized least squares (fGLS) estimator (Cameron & Trivedi, 2009). This technique is useful for cross-sectional time-series data because it estimates a GLS random-effects model with a weighted average of the between-subject (cross-section) and within-subject (fixed) effects (Lee, 2003). This is also a more conservative estimate in comparison with standard OLS or simple weighted least squares, which may result in downward-biased standard errors (Delmas et al., 2013). The standard errors are robust to within-panel heteroskedasticity as well as autocorrelation across time. Household effects are accounted for with fixed effects by unit.

The results reported in Table 1-2 (Model 1) show that treated households (i.e., households receiving reference group feedback) decreased their electricity usage from baseline overall ($\beta = -0.0017$, $z = -2.58$, $p = 0.010$), corresponding to a 1.5% decrease in consumption from baseline compared to control. Decomposing this effect by specific reference group, as shown in Model 2, the Average Reference Group treatment performed significantly better than control, leading to a greater decrease in 15-minute energy consumption from baseline compared to control ($\beta = -0.0076$, $z = -9.48$, $p < 0.001$). However, the Efficient Reference Group treatment performed significantly worse than control with *increases* in 15-minute energy consumption from baseline ($\beta = 0.0039$, $z = 5.13$, $p < 0.001$).[3] To put these findings in perspective, the treatment effect for the Average Reference Group corresponds to a 6.4% *decrease* in

[3] Note that some of the participants in this study were also participants in another study related to energy consumption the previous year conducted in the same residential facility. A regression that also includes a variable coding for participation in the prior study reveals a significant effect of prior participation but no substantive impact on the interpretation of the focal variables of this study. For full results, see Table A6 in the Appendix.

consumption from baseline compared to control, while the treatment effect for the Efficient Reference Group corresponds to a 3.7% *increase* in consumption from baseline compared to control. In practical terms, a 3.7% increase would represent an increase of 11.15 kWh/month, which equates to watching television for 110 hours (~4.5 days), working on a laptop for 223 hours (~9.3 days), or leaving on ten 100-Watt light bulbs for 11 hours. A 6.4% decrease would represent a decrease of 19.11 kWh/month, which equates to watching television for 190 hours (~8 days), working on a laptop for 382 hours (~16 days), or leaving on ten 100W light bulbs for 19 hours.

Table 1-2. Average treatment effects

	(1)	(2)	(3)
VARIABLES	15-minute kWh	15-minute kWh	15-minute kWh
Post treatment*Treated	-0.0017***		
	(0.0006)		
Post treatment*Average		-0.0076***	
		(0.0008)	
Post treatment*Efficient		0.0039***	
		(0.0008)	
Post treatment	-0.0093***	-0.0094***	
	(0.0013)	(0.0013)	
In a treatment group	0.0015***		
	(0.0006)		
Average reference group		0.0047***	
		(0.0007)	
Efficient reference group		-0.0013*	
		(0.0007)	
Children	0.0005**	0.0004*	0.0005***
	(0.0002)	(0.0002)	(0.0002)
Square feet	0.0001***	0.0001***	0.0001***
	(0.0000)	(0.0000)	(0.0000)
Floor	0.0097***	0.0094***	0.0097***
	(0.0002)	(0.0002)	(0.0002)
Env org member	-0.0147***	-0.0151***	-0.0145***
	(0.0005)	(0.0005)	(0.0005)
Constant	-0.0099***	-0.0108***	-0.0110***
	(0.0019)	(0.0019)	(0.0018)
Observations	1,297,780	1,297,780	1,297,780
Number of households	101	101	101

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01
Not reported: Controls for temperature, daylight savings, day of week, week in study, hour in day.

In order to test whether the average group performed significantly better than the efficient group, we estimated an additional model, this time using the Efficient Reference Group treatment—not the no-treatment control—as the reference group. This model confirmed that the Average Reference Group treatment performed significantly better than the Efficient Reference Group treatment (β = -0.0115, z = -13.11, $p < 0.001$). Thus, we can conclude from Study 1 that providing normative feedback lowered energy consumption overall, but that this drop is attributable exclusively to those compared to their average neighbors while those compared to their most efficient neighbors in fact increased energy consumption.

4.4. Target Effects: Heterogeneous Treatment Effects Depending on Characteristics

We next examine heterogenous responses to the Average and Efficient Reference Group treatments for (a) households with children and (b) those belonging to an environment group. We sought to assess whether these subgroups responded differently to each intervention compared to the rest of the sample. Recall that in Study 1, alongside the descriptive social comparison, participants also saw an injunctive social norm message regarding the negative consequences of energy consumption on both the environment and childhood asthma and cancer. We selected these messages based on pre-testing, finding that these two reasons for conservation were the most compelling to the building's residents. We were interested to see whether our main findings about normative feedback (i.e., that Average Reference Group comparisons cause consumption decreases while Efficient Reference Group comparisons cause consumption increases) might vary across critical individual differences between participants. Specifically, we wanted to see if the overall effect of normative comparisons was even stronger among environmentalists who were compared to the (more relevant) top-performing Target, and if the injunctive social norm regarding children's health was particularly effective for parents.

Given the mention of both children and the environment in the injunctive message, we expected this to hold particular meaning to households with children and those with particularly strong environmental concerns. We expected that those with relatively strong environmental concerns might be especially moved by comparisons to the Efficient Reference Group, and we expected that households with children would be especially moved by the injunctive message about children's health. To test these ideas, we looked at the interaction between our prior terms of interest (i.e., Post Treatment*Average, Post Treatment*Efficient) and key demographic variables: whether each household belonged to an environmental organization and whether each household had children. These results are reported in Table 1-3. Controlling for Distance and Valence (as we discuss in Sections 5.2 and 5.3; see Table 1-4), we can begin to explore the independent effects of different Target comparison groups above and beyond the effects of Distance and Valence.

The first model of Table 1-3 reveals that households with members who belong to an environmental organization did *not* reduce consumption from baseline more than control in response to the Average Reference Group treatment ($\beta = -0.0017$, $z = -0.94$, $p = .347$, *NS*). In contrast, households with members who belong to an environmental organization *did* reduce consumption from baseline in response to the Efficient Reference Group treatment compared to control ($\beta = -0.0042$, $z = -2.72$, $p = 0.007$). It seems as if the Efficient Reference Group may actually have been effective for those with a stronger environmental identity—even if it is not effective overall—but the Average Reference Group may not have been an effective target for environmentally conscious households. This provisional analysis suggests an important boundary condition: Normative comparisons may only be meaningful to people if the Target itself is personally meaningful. This is in line with previous work showing that group identification can moderate the effect of social comparisons for a given reference group (Göckeritz et al., 2010; Neighbors et al., 2010).

Table 1-3: Interactions with demographics.

	(1)	(2)
VARIABLES	15-minute kWh	15-minute kWh
Post treatment*Average*Children (binary)		-0.0105*** (0.0017)
Post treatment*Efficient* Children (binary)		0.0015 (0.0015)
Post treatment*Average*Env org member	-0.0337*** (0.0032)	
Post treatment*Efficient* Env org member	-0.0785*** (0.0029)	
Post treatment* Env org member	0.0198*** (0.0015)	
Average* Env org member	0.0335*** (0.0026)	
Efficient* Env org member	0.0798*** (0.0025)	
Post treatment*Average	-0.0045*** (0.0008)	-0.0050*** (0.0010)
Post treatment* Efficient	0.0097*** (0.0008)	0.0029*** (0.0009)
Post treatment* Children (binary)		-0.0025*** (0.0009)
Average*Children (binary)		-0.0120*** (0.0014)
Efficient*Children (binary)		-0.0348*** (0.0013)
Post treatment	-0.0116*** (0.0013)	-0.0079*** (0.0013)
Average reference group	0.0017** (0.0007)	0.0107*** (0.0009)
Efficient reference group	-0.0076*** (0.0007)	0.0107*** (0.0008)
Environmental organization member	-0.0354*** (0.0013)	-0.0125*** (0.0005)
Children (binary)		0.0163*** (0.0008)
Number of children	0.0003 (0.0002)	
Square feet	0.0001*** (0.0000)	0.0001*** (0.0000)
Floor	0.0091*** (0.0002)	0.0113*** (0.0002)
Constant	-0.0037* (0.0019)	-0.0230*** (0.0019)
Observations	1,297,780	1,297,780
Number of households	101	101

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01
Not reported: Controls for temperature, daylight savings, day of week, week in study, hour in day.

The second model in Table 1-3 reveals that households with children, in line with our main findings, reduced their consumption from baseline in response to the Average Reference Group treatment more than control (β = -0.017, z = -21.19, p < 0.001). However, contrary to our main findings, households with children *also reduced* their consumption from baseline in response to the Efficient Reference Group treatment compared to control (β = -0.027, z = 35.96, p < 0.001). The fact that households with children reduced consumption in response to both treatments suggests that the injunctive health message may have been more salient than the specific reference group for this segment of residents. Moreover, because this reduction from baseline is compared to a no-feedback control, this effect is not likely due simply to regression toward the mean.

While the Average Reference Group treatment in Study 1 was a more effective Target overall, these preliminary results suggest that this effect may be moderated by particular features of each household such as the presence of children or involvement in environmental organizations. Of course, because concern for the environment and having children was not randomly assigned, further research is needed to independently establish these findings. More generally, though, the significance of these moderators reveals that the particularities of the context of normative feedback may play an outsize role in determining its impact. When considering the independent impact of Target, one must also consider the values of the recipients of the feedback and their relationship to each of those Target groups—a topic for future research to address.

4.5. Discussion

Overall, we find a 2% decrease in electricity usage from providing normative feedback. However, this result is driven wholly by those who are compared to average neighbors—as opposed to the most efficient neighbors—who decreased their electricity usage by 6%. Those compared to the most efficient neighbors increased their electricity usage by about 4%. These findings are consistent with the claim that, *ceteris paribus*, people respond very differently to comparisons against average versus aspirational benchmarks. However, a second explanation of the results could be that the Valence of the feedback that most participants received was different between conditions. By definition, approximately 80% of participants in the Efficient Reference Group treatment underperformed the benchmark and thus received negatively valenced feedback. However, only about 50% of those in the Average Reference Group treatment were given negatively valenced feedback. It is possible that the pattern of results described above is at least partially driven by this Valence effect. As yet another explanation, perhaps these data resulted partially from the fact that, even among underperformers, more people were closer to the benchmark in the Average Reference Group treatment than the Efficient Reference Group treatment. For instance, someone in the 49th percentile of conservation would receive feedback that they were very *close* to the reference group if in the Average Reference Group treatment, but this same person would receive feedback that they were very *far* from the reference group if they were in the Efficient Reference Group treatment. Next, as best as these data will allow, we try to disentangle the distinct effects of Target, Valence, and Distance.

5. Decomposing Target, Valence, and Distance Effects on Household Energy Consumption

In order to better understand the mechanisms that drive the observed results, we estimated the impact of (a) the Valence of the feedback, and (b) the Distance to the reference group for each of the two treatments while controlling for Target. Although related to our randomly assigned Target treatments, both Valence and Distance may affect energy consumption in their own right. To test this possibility, we modeled Valence and Distance separately to isolate their effects from the independent effect of Target. In other words, we estimated the effects of Valence and Distance over and above the effect of Target. Then, controlling for these factors, we can also begin to further understand the independent contribution of Target.

It is important to treat the following results with due caution. Given the mechanical relationship between the (randomly assigned) Target and the resulting Valence and Distance feedback as described above, there are concerns of endogeneity. That is, because Valence and Distance were truthfully reported and not randomly assigned, participants' response to these separate factors may be related to their response to the randomly assigned Target. For corroborating evidence, we experimentally manipulate these factors in Studies 2 and 3. With that caveat, we offer the following analyses.

5.1. Valence Effects

To assess the effect of feedback Valence on energy consumption, we created weekly dummy variables for whether each treatment household received favorable (i.e., better than the reference group) or unfavorable (i.e., worse than the reference group) feedback. We found that 56% of weekly observations in the Average Reference Group treatment reflected favorable feedback, whereas 16% of weekly observations in the Efficient Reference Group treatment reflected favorable feedback. We used these variables to create cumulative feedback measures indicating (a) the cumulative proportion of favorable feedback messages that treatment households had received to date, and (b) the cumulative proportion of unfavorable feedback messages that treatment households had received to date. Finally, we used a negative exponential weighting function to weight more recent weeks of feedback more heavily than less recent weeks of feedback on the assumption that more recent feedback may be more salient and, thus, may have a larger impact on consumption behavior.

Our results, reported in Table 1-4 (Model 1) reveal that receiving favorable feedback is associated with significant *decreases* in consumption above and beyond the effect of assignment to Average or Efficient Reference Group treatment (β = -0.00025, z = -34.71, p < 0.001). This equates to a predicted 22% reduction from baseline relative to control for a household receiving exclusively favorable feedback throughout the entire treatment period. In contrast, receiving unfavorable feedback was associated with significant *increases* in consumption above and beyond the effect of assignment to Average or Efficient Reference Group treatment (β = 0.00017, z = 25.86, p < 0.001). This equates to a predicted 15% *increase* in consumption from baseline relative to control for a household receiving exclusively unfavorable feedback throughout the entire treatment period.

Table 1-4. Valence and distance effects

VARIABLES	(1) 15-minute kWh	(2) 15-minute kWh
Favorable valence	-0.0002***	
	(0.0000)	
Unfavorable valence	0.0002***	
	(0.0000)	
Favorable distance		-0.0007***
		(0.0000)
Unfavorable distance		0.0003***
		(0.0000)
Post treatment	-0.0106***	-0.0107***
	(0.0013)	(0.0013)
Average reference group	0.0051***	0.0060***
	(0.0006)	(0.0004)
Efficient reference group	-0.0060***	-0.0139***
	(0.0006)	(0.0004)
Children	0.0016***	0.0016***
	(0.0002)	(0.0002)
Square feet	0.0001***	0.0001***
	(0.0000)	(0.0000)
Floor	0.0085***	0.0068***
	(0.0002)	(0.0002)
Environmental organization member	-0.0148***	-0.0148***
	(0.0005)	(0.0005)
Constant	0.0047**	0.0349***
	(0.0019)	(0.0020)
Observations	1,146,345	1,146,345
Number of households	98	98

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Not reported: Controls for temperature, daylight savings, day of week, week in study, hour in day.

Overall, we see preliminary evidence that the Valence of feedback—independent of the specific Target or the Distance of that feedback—may play a significant role in shaping future behavior. The more positive feedback that people receive, the more likely they may be to continue cutting energy usage.

5.2. Distance Effects

We were interested not only in whether receiving favorable or unfavorable feedback had an impact on consumption but also in whether receiving *highly* favorable or highly unfavorable feedback (e.g., 200% better or worse than the reference group) impacts residents differently than receiving *slightly* favorable or slightly unfavorable feedback (e.g., 2% better or worse than the reference group). To this end, we examined the cumulative average feedback percentages that residents received (i.e., the cumulative average distance from the reference group of favorable and unfavorable feedback), again using a negative exponential weighting function to give more weight to more recent feedback.

Our estimation, reported in Table 1-4 (Model 2), reveals that energy consumption was significantly lower for larger distances of favorable feedback ($\beta = -0.00075$, $z = -45.10$, $p < 0.001$). This association corresponds to approximately a 7% reduction in consumption for every 10% better than the reference group that a household performed. In contrast, consumption was significantly higher for a larger distance of unfavorable feedback ($\beta = 0.00033$, $z = 56.68$, $p < 0.001$). This corresponds to approximately a 3% *increase* in consumption for every 10% worse than the reference group that a household performed. Thus, Distance may act as an amplifier—the directional response to favorable and unfavorable feedback is preserved, and the resulting change in consumption is magnified with a greater distance from the reference group. Because these factors were not randomly assigned, however, the true interaction between Distance and Valence is uncertain. For a test of the robustness of these results, see Appendix note N2.

5.3. Discussion

Taken together, our results suggest that peer comparisons may have a favorable impact when the feedback itself is favorable but a backfiring effect whenever this feedback is unfavorable. Moreover, this effect may be amplified by the Distance by which a person overperforms or underperforms relative to the reference group.

While Study 1 offered important clues about the effects of normative feedback, there are a few crucial limitations of the study. First, while Target was randomly assigned, the resulting Valence and Distance feedback was not randomly assigned. Instead, performance during week *W* was endogenous with feedback at the end of week *W* and thus potentially related to subsequent response in week *W+1*. This leaves open the concern that, for instance, people who are better at conservation respond differently to normative feedback regarding their consumption than those who are worse. A second limitation is that the selection of Target is not wholly orthogonal to the feedback regarding Distance and Valence. This means that we cannot fully disaggregate their independent contribution. Rather than being randomly assigned, Distance and Valence were mechanically determined by the random assignment to Target for a given level of conservation. Finally, we have estimated the cumulative effects of repeated treatment from normative Distance and Valence feedback, but not the independent effect of single-shot normative feedback. There are at least three concerns with this. First, there might be a difference between repeated versus one-off normative feedback. Second, we do not experimentally control—and thus cannot causally estimate—the effect of differences in the amount of exposure to treatment. Some participants checked their bills only monthly, whereas others voluntarily checked the web dashboard daily. Last, using a negative exponential weighting function, we have made assumptions about the effect of receiving mixed-valence feedback across many different exposures to normative comparisons (i.e., the effect of receiving positively valenced feedback some weeks and negatively valenced feedback other weeks), but these assumptions may not turn out to be correct. There could be order effects, dosage effects, contrast effects, or other interactions associated with multiple exposures to feedback that we could not observe in our analysis. Therefore, we randomly assign Distance and Valence feedback in Study 2 and Studies 3a-c, respectively, to address some of these limitations in Study 1.

6. Study 2: Daily Step Count and the Independent Effects of Distance

We believe one of the central reasons for the backfire effect associated with the Efficient Reference Group condition observed in Study 1 is that, by definition, (a) a far larger percentage of households received negatively valenced feedback, and (b) the households receiving negative feedback were at a greater Distance behind the reference group compared to those in the Average Reference Group condition. Study 2 was designed to further explore the effect of Distance in another domain of societal importance: personal health. In Study 2, we encourage people to walk more using normative feedback based on step-count data provided by a smartphone pedometer application. Study 2 seeks to address some of the limitations of Study 1. First, feedback was randomly assigned, which helps resolve some of the endogeneity concerns described above. Second, all participants were given feedback relative to the top performers (holding Target fixed) and were told that they underperformed compared to this benchmark (holding Valence fixed). Because Study 1 revealed backfiring effects of comparing households against top performers, in Study 2 we wanted to decompose the extent to which this was driven by the comparison against this specific Target group versus the extent to which it was driven by being a greater Distance from the benchmark. This design enabled us to cleanly establish the independent effect of Distance. Feedback was given only once, with a baseline observation period before the feedback was randomly assigned, and a treatment observation period after the feedback. The study therefore looks at the effects of a single instance of social comparison feedback, similar to an analysis of household energy consumption in the week immediately after the first round of feedback in Study 1 (see Appendix Table A1-5 for this analysis). Beyond this, our goal for Study 2 was to understand the mechanism driving the response to normative feedback observed in Study 1. Feeling relatively far from some benchmark has been shown to be demotivating (Rogers & Feller, 2016), especially when a distant comparison leads people to believe that their best effort will not be enough to match the benchmark (Lockwood & Kunda, 1999). We therefore hypothesized that informing participants that they were very far from some normative benchmark, compared to informing people that they were very close to the benchmark, leads to lowered subsequent performance because it lowers self-perceived ability to reach that benchmark. Hence, in Study 2 we measured participants' perceived self-efficacy to match or outperform the Target group in the post-feedback observation period.

6.1. Methods

In Study 2, based on power calculations and a pre-test to estimate attrition, we sought a final sample of 650 participants but, because of higher-than-expected attrition, we ended up with only 434 participants (M_{age} = 36.2, SD_{age} = 10.8, 60% female) recruited via Amazon's Mechanical Turk platform. Study 2 was conducted in three parts, with each part separated by a week in between: (1) initial recruitment, (2) random assignment to treatment, then (3) data upload and survey follow-up. Participants were paid at market rate for the platform as compensation for their time, and all participants were offered a $2 bonus if they completed all three parts. Initially we recruited 967 participants in part 1, but 306 participants (32%) failed to log in for part 2. Of the 661 participants who participated in part 2, an additional 227 (31%) failed to log in for Wave 3. A chi square test of independence finds that there was no significant differential attrition by condition (χ^2 (1, 208) = 0.012, p = 0.912).

In part 1, participants answered a set of screener questions to determine eligibility for participation in the study. These questions included whether they had an iPhone or Android smart phone, if they agreed to download an app and leave it on their phone for the duration of the study, and whether they had any physical limitations hindering their ability to walk. Participants who qualified were given instructions to download and set up a specific pedometer application, tested to ensure that the settings were correct, and verified that they were able to upload the app's data export. Finally, participants were reminded of the timeline and incentives of the study.

A week later, when participants logged in to participate in part 2, they uploaded their data and were briefly shown a "pinwheel" icon suggesting that calculations were being performed. Participants were then told, "Your performance was compared to the highest performers (the top 20%) of [participant's self-reported gender] aged [participant's age ± 5 years] who participated in a previous version of this study." Participants were randomly assigned to either the "near" or "far" condition and, depending on condition, told, "According to these analyses, you were 4 [39] percentage points behind the highest performers." The specific values for the number of percentage points behind the benchmark were decided upon after out-of-sample pre-testing in this setting.

Finally, in part 3, participants logged in once again and immediately uploaded their step-count data. Participants then answered some survey questions designed to assess the hypothesized psychological mechanism: motivation going into the next round (i.e., In Week 2, whether participants thought they could match the top performers from Week 1; -3 = definitely could not, +3 = definitely could). Finally, participants were debriefed on the false feedback that they received using both process and outcome debriefing (Ross et al., 1975). Analyses and methods were pre-registered and all materials and data for Study 2 can be found online.

6.2. Results and Discussion

Given the questions left open by Study 1, Study 2 was designed to isolate the effects of Distance by providing randomly assigned normative feedback. Moreover, Study 2 offered the chance to test the generalizability of the findings from Study 1 in an entirely new domain. Finally, most importantly, Study 2 enabled us to test for a psychological mechanism via survey measures that were administered in Study 2 but were not feasible to ask of participants in Study 1. We predicted that, compared to telling participants that they were close to the top-performers, telling participants that they were far from the top-performers would result in worse subsequent performance, as measured by the number of steps taken in the following week, mediated by their lowered perceived ability to match the Target group.

As anticipated, telling participants that they were relatively far from the benchmark resulted in substantially fewer steps the following week—an average of nearly 1,500 steps fewer—compared to those who were told they were relatively close. Analysis reveals that this total effect of condition, controlling for prior step-count, was considerable (β= -1497.14, t = -1.95, p = 0.052).[4] Our results further show that the relationship between Distance and subsequent performance was significantly mediated by perceived ability to match the Target group. As Figure 1-1 illustrates, the regression coefficient between Distance and perceived ability to match the reference group was statistically significant (β= -0.79, t =-5.09, p < 0.001), as was the regression coefficient relating perceived ability to match the reference group and subsequent

[4] Not having technical issues in our pre-test, we did not pre-register that analyses would exclude participants whose pedometer app reported them having taken 0 steps in Week 1 or Week 2, presumably the result of a technical glitch. Removing all participants whose data reported 0 steps in either wave, as well as one participant whose data reported only having taken 15 steps all week, left a sample with no one taking fewer than 100 steps in the week. Estimating the effects after dropping these three participants yielded a similar, significant effect of condition on increased step-count (β = 1514.91, t = -1.96, p = 0.050).

performance (β = 824.45, t = 3.37, p < 0.001). The estimate of the indirect effect was -652.98 steps.

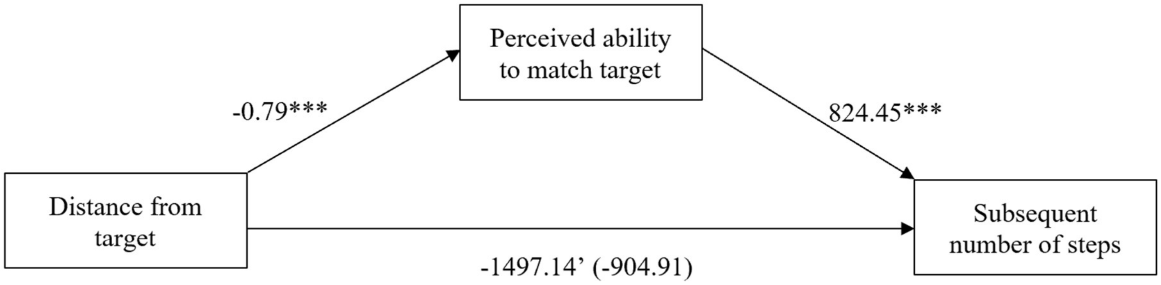


Fig. 1-1 Regression coefficients for the relationships between distance and subsequent steps as mediated by perceived ability to match the target. The coefficient of the direct effect of distance on subsequent number of steps walked, controlling for perceived ability to match, is in parentheses.
'p < .06, *p < .05, **p < .01, ***p < .001.

We tested the significance of this indirect effect using bootstrapping procedures. Unstandardized indirect effects were computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was estimated. The bootstrapped unstandardized indirect effect of -652.98 had a 95% confidence interval of [-1190.31, -235.62], thus the indirect effect through the perceived ability to match Target was statistically significant. The estimated average treatment effect suggests that the indirect effect of being relatively far from the benchmark via its negative impact on perceived ability to match the Target is associated with taking about 650 steps fewer the following week. Considering these results alongside the findings from Study 1, we conclude that receiving negative social comparison information reduces people's sense of their ability to match the Target benchmark, which in turn reduces their post-feedback performance. This finding is consistent with the pattern reported by Rogers and Feller (2016), who found that students who observed the work of very high-performing peers felt that this level of achievement was not personally attainable and thus were more likely to perform worse in an online course.

7. Study 3: Three Experiments Testing the Independent Effects of Valence

As discussed in Section 2.3, it is an open question how the Valence of feedback might affect performance. On the one hand, our provisional decomposition results of Study 1, in line with others (e.g., Tybout & Yalch, 1980), showed that positively valenced feedback engenders higher subsequent performance than negatively valenced feedback. On the other hand, however, past research (e.g., Schultz et al., 2007) has shown that negatively valenced feedback can cause performance *improvement* and, at least in the absence of an injunctive norm, positively valenced feedback can cause performance *reductions*. Relatedly, Berger and Pope (2011) find that basketball teams who are losing by one point at halftime are more likely to win than the team who is ahead. Together, the experiments in Study 3 sought to determine the impact of the Valence of normative feedback while holding constant the effects of Distance and Target. Moreover, we sought to investigate this phenomenon in additional domains of task performance. Our goal for Study 3a was to investigate how Valence impacts motivation toward future performance in a word-search task. Then, in a design similar to Study 2, in Study 3b we tested an indirect effect of Valence on step-count through the measures of motivation used in Study 3a. Finally, in Study 3c we tested how receiving negatively valenced feedback on a typing test impacts performance compared to receiving positively valenced feedback. Analyses and methods were all pre-registered, and all materials can be found online.

7.1 Study 3a: Word Search and the Motivational Impact of Valence

After receiving normative feedback, people make meaning of this feedback in ways that affect their future performance. We were interested in determining whether negatively valenced feedback (compared to positively valenced feedback) is dispiriting, causing people to feel demotivated and demoralized, or if it is instead encouraging, inspiring feelings of motivation to strive to match the Target. Thus, in Study 3a we randomly assigned positively or negatively valenced feedback, holding Distance and Target constant, and measured participants' reactions to the feedback and plans for subsequent effort toward a similar task.

7.1.1. Study 3a Methods

In Study 3a, we recruited a sample of 436 participants from Amazon's Mechanical Turk platform who passed our attention checks. After excluding participants based on our pre-registered exclusion criteria—those who failed a manipulation check asking them to recall the Valence of their feedback and those who reported not believing the feedback in an open-response question—we were left with 305 participants (M_{age} = 38.8, SD_{age} = 12.2, 50% female). There was no significant differential attrition by condition for failure on the Valence manipulation check (χ^2 (1, 436) = 0.659, p = 0.417) nor for failure on the open-response mentions of false feedback (χ^2 (1, 436) = 0.108, p = 0.742). After an initial attention screener, participants were informed that they would participate in two rounds of a word-search task in which their job was to find as many words as possible in a 20 x 30 array of letters. They were informed that they would receive peer-comparison feedback in between rounds, and then they began the Round 1 word-search task lasting three minutes. At the end of Round 1, participants were (truthfully) told the number of words they found and were randomly assigned to be informed that they performed either 59% better or 59% worse than the average of the top 20 participants who had previously completed the word-search task.

After receiving this feedback, participants were asked a series of three questions designed to understand how the feedback impacted their motivation going into Round 2. Participants were asked, in a random order, how discouraged or encouraged they felt about their word-search abilities (-4 = Extremely discouraged, +4 = Extremely encouraged), how motivated or demotivated they were feeling going into Round 2 (-4 = Extremely demotivated, +4 = Extremely motivated), and how well they expected to do in Round 2 (1 = Terrible, 5 = Excellent). This final question was rescaled in our analysis to match the range of the other two measures.

Per our pre-registration plan, Cronbach's alpha was calculated for the three questions that participants were asked after receiving the (false) feedback. This test showed an acceptable reliability of the three items ($\alpha = 0.85$), with no improvements to reliability resulting from excluding any of the measures. Thus, an index of motivation was constructed by averaging the scores on the three items for each participant, forming the key dependent measure of interest in Study 3a.

7.1.2. Study 3a Results and Discussion

The central purpose of Study 3a was to explore how the Valence of feedback impacted participants' motivation going into the next round of the task. We wanted to test whether negative feedback was discouraging or encouraging of planned effort for subsequent performance. Of course, planned effort is a noisy measure of actual effort expended in later rounds, and performance on this task is itself only partially related to pure effort (i.e., simply trying harder may not be enough to actually find more words in a word-search task just as wanting to conserve more energy may not actually result in greater conservation). Nonetheless, how feelings, motivation, and planned effort shift in response to positive versus negative feedback is an important first insight to understanding the effects of Valence.

On a 9-point motivation scale centered at zero, the average score for participants who were told that they performed better than the Target was 1.83 ($SD_{positive}$=1.14). However, as predicted, telling participants that they performed worse than the Target resulted in significantly lower motivation (M_{neg}=-0.46, SD_{neg}=1.53; β = -2.29, t = -14.81, p<.001). Because this feedback was randomly assigned, we take this as evidence that receiving negative social comparison information—at least when comparing a Distance of 59% ahead of or behind a Target of top performers—reduces people's sense of their abilities and motivation to perform well in later rounds. Of note, we do not find evidence that the impact of negatively valenced feedback varies systematically with absolute performance during Round 1 ($\beta_{\text{interaction}} = -0.12$, $t = -0.39$, p=0.700). Thus, it seems, *ceteris paribus*, that positively valenced feedback leads to higher motivation compared to negatively valenced feedback.

7.2 Study 3b: Daily Step Count and the Effect of Valence

Study 3a demonstrated that positively valenced feedback engenders greater motivation compared to negatively valenced feedback. In Study 3b we wanted to see whether this motivating effect of positive Valence would in turn result in more improved performance compared to negatively valenced feedback. To study this, we replicated the crucial elements of Study 2's design using a pedometer smartphone application, this time testing the independent effect of Valence.

7.2.1. Study 3b Methods

Study 3b was a replication of Study 2 with a few critical changes. First, each wave of observation lasted only one day rather than six. Second, the feedback that was randomly assigned to participants between Wave 1 and Wave 2 manipulated the Valence of the normative comparison, not the Distance. After uploading their Wave 1 data, participants were told that they performed either 39% better or 39% worse than the highest performers participating in the study. The top-performing group was once again defined as the top 20% of participants, but this time we did not indicate that participants were only being compared to members of their age and gender bracket.

In Study 3b, based on the assumption that attrition rates and the effect size would be similar to those observed in Study 2, we sought to recruit a sample of 1500 participants to the initial screener with a goal of a final sample of about 800 participants. Unfortunately, we were only able to recruit 1069 participants to the initial screener. Of them, only 195 qualified to participate in the study based on possessing either an iPhone or Android-based smartphone, willingness to download the pedometer app and share data with us, self-report of having minimal to no physical limitations on their ability to walk, and passing our attention checks. Of them, 72 participants (37%, similar to Study 2) did not log in for the part 2 data upload and experimental manipulation of feedback. Of the 123 participants who uploaded their data and received feedback at this mid-point (the experimental manipulation), only two did not log in two days later for the final data upload. In the end, we had a final sample of 121 participants (M_{age} = 36.1, SD_{age} = 10.6, 41% female) of Amazon Mechanical Turk workers. Of them, 61 participants were randomly assigned to receive negatively valenced feedback and the other 60 participants received positively valenced feedback.

There were two critical measures of interest in Study 3b. First, we measured Wave 2 step-count after receiving the randomly assigned Valence feedback. Second, we constructed a "Wave 2 Motivation" index similar to the dependent measure used in Study 3a. We wanted to see if this Wave 2 Motivation index mediated the effect of feedback Valence on Wave 2 step-count.

7.2.2. Study 3b Results and Discussion

Our results show, as predicted, a significant indirect effect of Valence on subsequent step-count through the Wave 2 Motivation index. Receiving positive feedback significantly increased Wave 2 motivation (i.e., the a-path) compared to negative feedback ($M_{negative}$=-0.35,

$M_{positive}$=2.57; β = 2.92, t = 7. 24, p<.001). The association between motivation and Wave 2 step-count while controlling for Wave 1 step-count (i.e., the b-path) was nearly statistically significant (β = 205.4, t = 1.97, p=0.051). Together, the estimated indirect effect was 584.22 steps. We tested the significance of this indirect effect again using 10,000 bootstrapped samples to compute the 95% confidence interval. The bootstrapped indirect effect had a 95% confidence interval of [47.66, 1156.74], thus the effect of feedback Valence on subsequent step-count was significantly mediated by the Wave 2 Motivation index. In sum, we find evidence consistent with the hypothesis that, compared to negatively valenced feedback, positively valenced feedback encourages motivation and results in improved performance.

We note that we did not observe a statistically significant total (unmediated) effect on step-count the day after receiving feedback. The average number of steps during the one-day observation period increased from Wave 1 to Wave 2 for both groups, with those receiving positively valenced feedback (M_{growth} = 426.3, SD_{growth} = 2839.2) improving by over three times as many steps as those receiving negatively valenced feedback (M_{growth} = 122.3, SD_{growth} = 2472.3). Nonetheless, this difference of more than 300 steps did not approach statistical significance (β = 382.7, t = 0.82, p=0.41), perhaps due to the drastically smaller-than-anticipated sample size and extremely large standard errors that resulted.

7.3 Study 3c: Feedback Valence and Typing Speed

The goal of Study 3c was to weave together our findings from Study 3a and Study 3b to establish an effect of feedback Valence on subsequent performance. To do this, we moved into the domain of typing performance, one that is largely though not wholly influenced by effort. That is, while some amount of typing performance is determined by skill and practice, some degree of typing performance can be explained simply in terms of in-the-moment effort and care.

Moreover, there is an implicit injunctive social norm of "more is better." For this reason, we tested the effect of randomly assigned Valence of normative feedback on typing performance. Further, to test the generalizability of the effect of Valence, in Study 3c we randomly assigned positively or negatively valenced feedback compared to a Target of average performers (unlike Studies 2, 3a, and 3b, in which we used a Target of top performers).

7.3.1. Study 3c Methods

In Study 3c, we recruited a sample of 597 participants from Prolific who passed our attention checks (M_{age} = 34.5, SD_{age} = 12.3, 46% female). As pre-registered, we also conducted all analyses described below after excluding participants who indicated that they did not believe that the feedback they received between Round 1 and Round 2 was an accurate report of their performance compared to others. This left a sample of 439 participants (M_{age} = 33.9, SD_{age} = 12.0, 46% female). It is noteworthy that, unlike in other studies, there were different rates of disbelieving the veracity of the feedback between conditions with this study design (χ^2 (1, 597) = 19.661, p <.001).

Study 3c took place in three critical phases. In part 1, after an introduction, participants were informed that they would have 30 seconds to transcribe as much of a block of text as they could. Participants were told to move as quickly as possible but informed that they would be judged based only on the number of *correctly* typed words, so performance was measured as accurate typing. Then in part 2, after the first typing test, participants were informed of their (true) number of words typed in Round 1, and were randomly assigned to be told that they had performed either 39% better or 39% worse than the average participants in the study. They were then told, "Remember to type quickly, but your words only count if there are no typos." Finally, in part 3, participants had six minutes to type as much as they could as accurately as they could.

We made Round 2 considerably longer than Round 1 because we speculated that any differences in motivation caused by the Valence of feedback would be amplified if participants were given a longer time in Round 2.

Our critical dependent measures of interest were captured in Round 2: accuracy and total wordcount. First, we counted the total number of words typed during Round 2 of the typing test. Then, we used these words that participants typed to construct a measure of typing performance: how accurately participants transcribed the text. We wanted to see if those who were given positively valenced feedback would be more accurate in the text that they transcribed than those who were given negatively valenced feedback. To measure this, we calculated a similarity score using the Optimal String Alignment ("OSA") method, a method similar to the Damerau-Levenshtein Distance calculation. The OSA compares the text that was inputted with the part of the original text that was transcribed, allowing for transposition of adjacent characters. These scores, based on the number of changes that would be required to make the inputted text match the original, represent the similarity of the inputted text to the original. OSA scores range from 0 to 1, where a score of 1 means that the inputted text perfectly matches the original text. We chose to measure these two dimensions of speed and accuracy separately in order to both have a finer resolution for detecting differences on each individual dimension, and also to separately capture the effects for these two dimensions on which participants could trade off.

7.3.2. Study 3c Results and Discussion

In Study 3c, we were interested in the effect of feedback Valence on typing test performance. First, we find no significant effect when using the full sample of participants. Including those who rejected the veracity of our feedback—a full 26% of participants—was enough to eliminate any effect of condition.[5] Turning toward the pre-registered analysis of those participants who did not report disbelieving the accuracy of the feedback, we find the predicted positive effect of positively valenced feedback. In the typing test, positively valenced feedback caused an improvement in typing performance during Round 2 ($\beta = 0.018$, t = 2.20, $p<.05$). This relationship is similar even when controlling for Round 1 performance ($\beta = 0.015$, t = 2.14, $p<.05$).

In addition to examining the effect of positively or negatively valenced feedback on typing performance, we also examined whether valence of feedback had an impact on the total number of words that participants typed. As with typing performance, the number of total words typed during Round 2 increased from receiving positively valenced feedback ($\beta = 29.8$, t = 10.01, $p<.004$). However, this relationship was no longer statistically significant once controlling for the number of words typed during Round 1.[6]

[5] As one might expect, participants who did not believe our feedback were less impacted by it. For those receiving negative feedback, participants who did not believe the negative feedback were more motivated and performed better than those who did believe the negative feedback (Motivation: M_{believed}=0.8 vs. $M_{\text{disbelieved}}$=1.1; Performance: M_{believed}=0.95 vs. $M_{\text{disbelieved}}$=0.97; Wordcount: M_{believed}=237 vs. $M_{\text{disbelieved}}$=312). The opposite pattern held for those who received positive feedback and did not believe it: Participants who did not believe the positive feedback were less motivated and performed worse than those who did believe the positive feedback (Motivation: M_{believed}= 1.8 vs. $M_{\text{disbelieved}}$=1.1; Performance: M_{believed}=0.97 vs. $M_{\text{disbelieved}}$=0.95; Wordcount: M_{believed}=267 vs. $M_{\text{disbelieved}}$=247).

[6] Perhaps explaining this, a model predicting Round 2 wordcount just from Round 1 wordcount alone shows that 73% of variation in Round 2 wordcount can be explained simply from Round 1 wordcount ($\beta = 9.87$, t = 34.4, $p<.001$). However, this relationship is much weaker for overall typing performance, with Round 1 accuracy explaining just 10% of variation in Round 2 accuracy ($\beta = 0.45$, t = 7.15, $p<.001$). We did not anticipate Round 1 wordcount impacting Round 2 wordcount so much. In an analysis that was not pre-registered, we calculated a different dependent measure by subtracting Round 1 wordcount from Round 2 wordcount. This "difference score" effectively takes Round 1 performance into account while constraining the amount of variation it is able to explain in the model. On this specification, Valence is once again statistically significant, with positively valenced feedback leading to more improvement ($\beta = 27.21$, t = 2.91, $p<.004$).

Table 1-5a. Study 3c Results: Accuracy

	Dependent variable:	
	Round 2 Accuracy	
positive valence	0.018^{*}	0.015^{*}
	(0.008)	(0.007)
r1 accuracy		0.457^{***}
		(0.063)
intercept	0.950^{***}	0.521^{***}
	(0.006)	(0.060)
Observations	439	435
R^2	0.011	0.115
Adjusted R^2	0.009	0.111
F Statistic	4.829^{*} (df = 1; 437)	28.062^{***} (df = 2; 432)
Note:		*p<0.05 **p<0.01 ***p<0.001
		The 'positive valence' condition variable is dummy-coded

Table 1-5b. Study 3c Results: Wordcount

	Dependent variable:			
	Round 2 Wordcount			Growth Score
positive valence	29.803^{**}	4.297		27.210^{**}
	(10.077)	(5.339)		(9.345)
r1 wordcount		9.836^{***}	9.869^{***}	
		(0.290)	(0.287)	
intercept	237.144^{***}	17.908^{*}	19.527^{**}	214.856^{***}
	(7.528)	(7.571)	(7.296)	(6.981)
Observations	439	439	439	439
R^2	0.020	0.731	0.730	0.019
Adjusted R^2	0.017	0.730	0.730	0.017
F Statistic	8.747^{**} (df = 1; 437)	591.848^{***} (df = 2; 436)	$1,184.003^{***}$ (df = 1; 437)	8.478^{**} (df = 1; 437)
Note:				*p<0.05 **p<0.01 ***p<0.001
				The 'positive valence' condition variable is dummy-coded

These results, when joined with Study 3a, Study 3b, and the decomposition analyses of Study 1, demonstrate the independent effect of the Valence of normative comparisons. While normative comparison feedback may in aggregate improve performance, we show that this effect is not uniform across those receiving favorable versus unfavorable comparisons.

8. General Discussion

Taken together, the results of Studies 1 - 3 begin to paint a picture of the independent contribution of Target, Distance, and Valence in explaining people's response to normative feedback. Overall, we find in Study 1 that receiving peer comparison feedback resulted in decreases in household energy consumption (-2%). However, while our Average Reference Group treatment resulted in large energy consumption *decreases* (-6%), our Efficient Reference Group treatment resulted in energy consumption *increases* (+4%). As a first step to investigating why this might be the case, we found that receiving more negatively valenced feedback was associated with an increase in consumption (+15%), while receiving more positively valenced feedback was associated with decreased consumption (-22%). However, we found that the effect of different Targets depended on household-specific characteristics. Moreover, decomposition analyses showed that the Distance was correlated with the size of the response. Then, using randomized (false) feedback to establish the independent effect of Distance, in Study 2 we found that being further from a benchmark decreases subsequent performance through a lowered belief in ability to match that Target. Finally, in Studies 3a-c, using randomized feedback to identify the independent effect of feedback Valence, we find that negatively valenced feedback lowers motivation and subsequent effort, which in turn is associated with worse subsequent performance. Our chief purpose in this project was to establish these three dimensions as critical elements of peer comparison feedback.

8.1. Target, Distance, and Valence: The Three Elements of Peer Comparison Feedback

When considering the effects of peer comparison feedback, it is important to account for how the individuals receiving the feedback perform relative to the reference group. In the current investigation, we find that peer comparison feedback is not uniformly helpful but depends on the specific Target, Distance, and Valence for each feedback recipient. Moreover, we find that individual demographic and attitudinal factors—in this case, whether a person has children or strong pro-environmental beliefs—may moderate the effect of Target. This suggests that whenever choice architects intent on providing normative feedback have the option to select a reference group, they should do so carefully. In choosing a reference group, the choice architect is also choosing to selectively motivate some and demotivate others based on the proportion of people who will receive negatively valenced feedback and how far from the Target people will be on average.

8.2. Future Directions

Having the framework of three separate dimensions of Target, Distance, and Valence allows for a deeper, more nuanced future study of normative feedback. Using these independent elements of peer comparison, we now turn toward insights and questions raised from our investigation to motivate future research.

One avenue for further exploration relates to how different labels for a given reference group (i.e., Target) can influence behavior. Whereas Meeker and colleagues (2016) find an overall positive effect on lowering inappropriate antibiotic prescribing rates by comparing doctors against “top-performers,” we find an overall *negative* impact of comparing neighbors against their “most efficient” peers. One factor that could be driving this divergent pattern of results is a different attitude toward the labels themselves among the different populations. For instance, how doctors feel about being compared to top performers may differ from how homeowners feel about being compared to their most efficient neighbors.

Additionally, when people have an internal reference point (Bell & Bucklin, 1999) of the level of performance that is judged to be acceptable (e.g., walking 10,000 steps or consuming 2,000 calories per day as a meaningful benchmark), this may moderate the effect of normative feedback. When performance standards are more uncertain or harder to evaluate, people may look to peers' performance more when evaluating personal performance. Additionally, it is possible that there are individual differences (e.g., growth mindset) that may explain why some people respond positively to positive feedback and others are more motivated by negative feedback.

Further, more work is needed to identify differences and similarities in the effects of Target, Distance, and Valence of single-shot versus repeated normative messaging, as well as the effect of varying frequency of messages. For instance, consider someone who receives negatively valenced feedback after Wave 1 who then, in response, increases effort. Whether that person receives no more feedback, negative feedback, or positive feedback after Wave 2 would likely have a sizeable impact on continued motivation, effort, and performance. This sort of experimentation and learning from feedback often, though not always, occurs when people receive normative comparisons. Moreover, if feedback is repeated, the frequency and timing of feedback could exert substantial influence on performance and whether an identity as a top- or low-performer begins to set in. Someone who constantly receives similar feedback may quickly develop a stable identity as an under- or top-performer, leading to long-run behavior change. If, however, frequent feedback typically vacillates between positively and negatively valenced messaging, this could have an overall backfiring effect if it undermines a person's feelings of self-efficacy and control (Major et al., 1991).

Last, future work should consider a fully crossed design of randomized Target, Distance, and Valence social comparisons in order to explore the possibility of various two- or three-way interactions. We expect that the three independent effects of Target, Distance, and Valence do in fact interact to shape people's responses. It is conceivable that the effect that Valence has on Distance may change if people are compared to an average versus a top-performing Target. Perhaps, for instance, being a small distance behind a top-performing target is more motivating than being a medium distance ahead of a low-performing target. It is important, therefore, to separate and interact all three of these dimensions independently in a controlled way, accounting for the moderators discussed throughout. If Target, Distance, and Valence all interact in their combined effect on people, disentangling these forces may help to explain sometimes differing results observed in the literature (e.g., Meeker et al., 2016 versus Schultz et al., 2007).

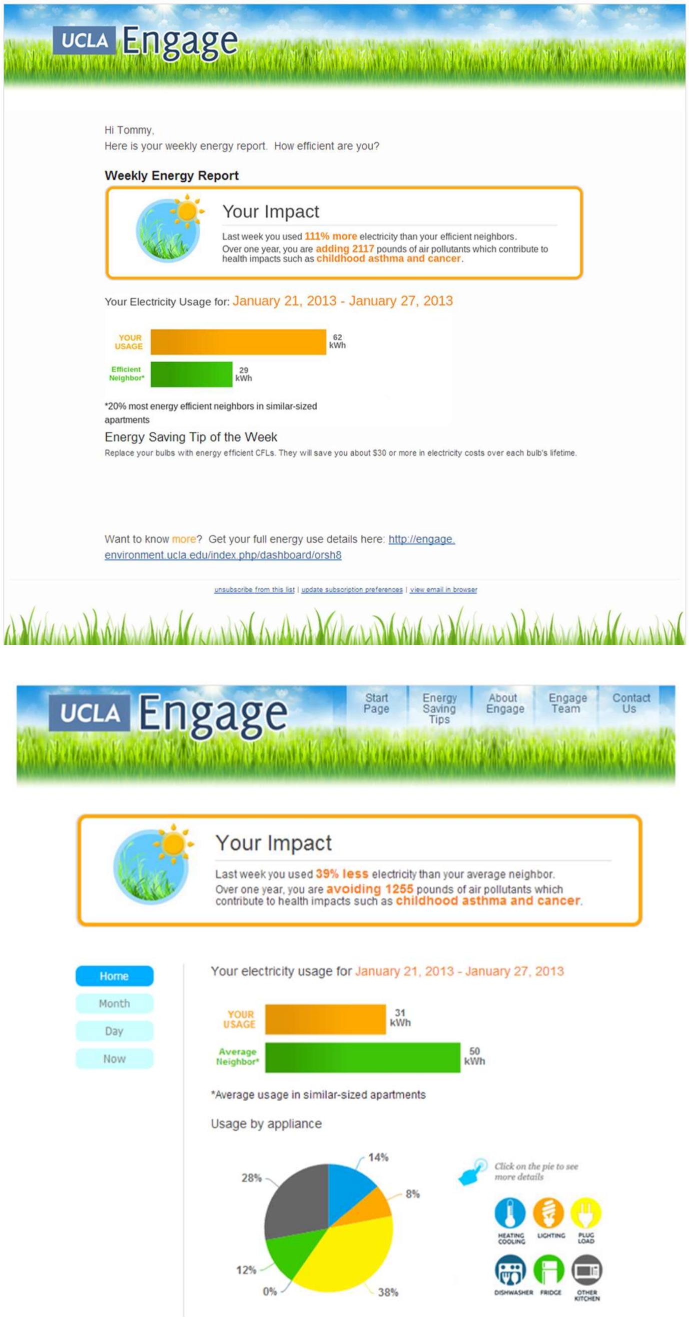
9. Conclusion

Any time a teacher posts the results of a class test or a manager posts a quarterly performance report, they are giving people the chance to compare their performance against others. This feedback can cause sizeable changes in subsequent behavior depending on who the Target group is, how far one is from that benchmark, and whether one overperformed or underperformed against that benchmark. Moreover, each of these factors can have a very different effect on motivation and performance depending on a host of other moderating variables. Choosing to compare people against some normative benchmark necessarily means choosing a Target, Distance, and Valence of feedback. The would-be choice architect must therefore make these decisions wisely. Absent careful thinking, social comparison feedback can backfire significantly. Thus, choice architects might instead consider implementing individually tailored social comparisons whereby different people, depending on what they care about and what their baseline performance is, are compared against different benchmarks. Harnessed wisely, normative feedback can improve health, wellbeing, performance, and behavior toward the common good.

Appendix

Fig. A1-1. Messages shown to participants in the treatment conditions.

A



Hi Tommy,
Here is your weekly energy report. How efficient are you?

Weekly Energy Report

Your Impact

Last week you used 111% more electricity than your efficient neighbors.
Over one year, you are adding 2117 pounds of air pollutants which contribute to health impacts such as childhood asthma and cancer.

Your Electricity Usage for: January 21, 2013 - January 27, 2013

*20% most energy efficient neighbors in similar-sized apartments

Energy Saving Tip of the Week

Replace your bulbs with energy efficient CFLs. They will save you about $30 or more in electricity costs over each bulb's lifetime.

Want to know more? Get your full energy use details here: http://engage.environment.ucla.edu/index.php/dashboard/orsh8

unsubscribe from this list | update subscription preferences | view email in browser

B

Your Impact

Last week you used 39% less electricity than your average neighbor.
Over one year, you are avoiding 1255 pounds of air pollutants which contribute to health impacts such as childhood asthma and cancer.

Your electricity usage for January 21, 2013 - January 27, 2013

*Average usage in similar-sized apartments

Usage by appliance

Figure A1-1. *(A) Example weekly email that a participant in the Efficient Reference Group treatment might have received. (B) Example dashboard that a participant in the Average Reference Group treatment might have seen on the website.*

Table A1-1. Per capita residential energy consumption

Region	2010 Population (in thousands)	Annualized kWh	kWh per capita
United States*	308,746	3,749,985 x 10^6	12,146
California*	37,254	250,384 x 10^6	6,721
LADWP*	1400	8017.65 x 10^6	5,726
Graduate Housing	0.518	2910.782	5,619

*Source: California Energy Commission data, 2010

Table A1-2. Emails opened

VARIABLES	(1) Controlling for whether saw email/dashboard	(2) Excluding households that didn't see email/dashboard
Post treatment*Average	-0.0070***	-0.0035***
	(0.0008)	(0.0008)
Post treatment*Efficient	0.0046***	0.0049***
	(0.0008)	(0.0008)
Post treatment	-0.0097***	-0.0095***
	(0.0013)	(0.0013)
Average reference group	0.0466***	-0.0027***
	(0.0011)	(0.0007)
Efficient reference group	0.0410***	-0.0020***
	(0.0011)	(0.0007)
Saw dashboard/email	-0.0452***	
	(0.0009)	
Children	0.0014***	0.0011***
	(0.0002)	(0.0002)
Square feet	0.0001***	0.0001***
	(0.0000)	(0.0000)
Floor	0.0089***	0.0081***
	(0.0002)	(0.0002)
Environmental org member	-0.0138***	-0.0143***
	(0.0005)	(0.0005)
Constant	-0.0086***	-0.0112***
	(0.0019)	(0.0019)
Observations	1,297,780	1,247,964
Number of households	101	96

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A1-3. Robustness checks for distance and valence effects

	(1)	(2)
VARIABLES	**Hourly kWh**	**Hourly kWh**
Favorable valence	-0.0009***	
	(0.0002)	
Unfavorable valence	0.0005**	
	(0.0002)	
Favorable distance		-0.0012*
		(0.0005)
Unfavorable distance		0.0011***
		(0.0001)
Post treatment	-0.1845***	-0.1839***
	(0.0298)	(0.0293)
Avg reference group	-0.1122	-0.0711
	(0.0577)	(0.0566)
Efficient reference group	0.0607	0.0864
	(0.0611)	(0.0603)
Children	-0.1726***	-0.1872***
	(0.0319)	(0.0321)
Square feet	0.0037***	0.0039***
	(0.0003)	(0.0003)
Floor	0.1250***	0.1326***
	(0.0325)	(0.0324)
Env org member	-0.1586	-0.1989
	(.01021)	(0.1031)
Constant	-2.9900***	-3.1570***
	(0.2761)	(0.2775)
Observations	286,276	286,276
Number of households	97	97

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A1-4. Impact of environmental factors

VARIABLES	(1) Hot weather	(2) Cold weather	(3) Daytime	(4) Nighttime	(5) Weekdays	(6) Weekends
Post treat*Avg	-0.0009	-0.0091***	-0.0145***	-0.0002	-0.0063***	-0.0092***
	(0.0020)	(0.0012)	(0.0012)	(0.0010)	(0.0009)	(0.0016)
Post treat*Eff	0.0164***	0.0006	0.0080***	-0.0008	0.0040***	0.0031**
	(0.0020)	(0.0012)	(0.0010)	(0.0010)	(0.0008)	(0.0015)
Post treat	-0.0081**	-0.0052***	-0.0080***	-0.0110***	-0.0087***	-0.0159***
	(0.0040)	(0.0014)	(0.0017)	(0.0016)	(0.0013)	(0.0019)
Avg ref group	0.0018*	0.0052***	0.0140***	-0.0042***	0.0038***	0.0060***
	(0.0010)	(0.0012)	(0.0010)	(0.0008)	(0.0008)	(0.0014)
Eff ref group	-0.0017	-0.0024**	-0.0055***	0.0012	-0.0020***	0.0006
	(0.0011)	(0.0011)	(0.0009)	(0.0009)	(0.0008)	(0.0013)
Children	-0.0020***	0.0020***	-0.0000	-0.0005**	0.0001	0.0008**
	(0.0005)	(0.0002)	(0.0003)	(0.0002)	(0.0002)	(0.0004)
Square feet	0.0002***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Floor	0.0173***	0.0074***	0.0090***	0.0103***	0.0089***	0.0104***
	(0.0005)	(0.0002)	(0.0003)	(0.0002)	(0.0002)	(0.0004)
Env org member	-0.0087***	-0.0164***	-0.0181***	-0.0112***	-0.0150***	-0.0126***
	(0.0015)	(0.0005)	(0.0006)	(0.0008)	(0.0006)	(0.0011)
Constant	-0.1077***	0.0141***	-0.0110***	-0.0064***	-0.0116***	-0.0170***
	(0.0052)	(0.0019)	(0.0026)	(0.0024)	(0.0020)	(0.0033)
Observations	336,121	913,605	648,914	648,866	918,415	379,365
Number of households	99	101	101	99	100	100

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Not reported: Controls for temperature, daylight savings, day of week, week in study, hour in day.

Table A1-5. Effects across time

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Var	1 wk	2 wks	3 wks	4 wks	5 wks	6 wks	7 wks	8 wks	9 wks	10 wks	11 wks	12 wks
Post treat*Avg	0.0033	-0.0038**	-0.0058***	-0.0071***	-0.0067***	-0.0079***	-0.0092***	-0.0108***	-0.0110***	-0.0109***	-0.0093***	-0.0093***
	(0.0025)	(0.0018)	(0.0015)	(0.0014)	(0.0012)	(0.0012)	(0.0011)	(0.0011)	(0.0010)	(0.0010)	(0.0010)	(0.0010)
Post treat*Eff	0.0072***	0.0090***	0.0082***	0.0075***	0.0074***	0.0064***	0.0044***	0.0026**	0.0016	0.0015	0.0020**	0.0020**
	(0.0023)	(0.0017)	(0.0015)	(0.0013)	(0.0012)	(0.0011)	(0.0010)	(0.0010)	(0.0010)	(0.0010)	(0.0010)	(0.0010)
Post treat	-0.0088***	-0.0093***	-0.0088***	-0.0089***	-0.0091***	-0.0086***	-0.0085***	-0.0075***	-0.0074***	-0.0073***	-0.0085***	-0.0088***
	(0.0018)	(0.0016)	(0.0015)	(0.0014)	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0012)
Avg ref group	0.0059***	0.0064***	0.0068***	0.0071***	0.0071***	0.0075***	0.0069***	0.0066***	0.0054***	0.0049***	0.0036***	0.0038***
	(0.0009)	(0.0009)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Eff ref grp	-0.0043***	-0.0036***	-0.0033***	-0.0036***	-0.0034***	-0.0023***	-0.0018**	-0.0019**	-0.0032***	-0.0028***	-0.0035***	-0.0036***
	(0.0009)	(0.0009)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Children	0.0034***	0.0022***	0.0008**	0.0006*	-0.0003	-0.0010***	-0.0009***	-0.0006**	0.0007**	0.0013***	0.0015***	0.0015***
	(0.0005)	(0.0004)	(0.0004)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Square feet	0.0002***	0.0002***	0.0002***	0.0002***	0.0002***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Floor	0.0186***	0.0172***	0.0162***	0.0160***	0.0152***	0.0142***	0.0141***	0.0146***	0.0153***	0.0154***	0.0155***	0.0151***
	(0.0004)	(0.0004)	(0.0004)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Env org member	-0.0047***	-0.0099***	-0.0122***	-0.0135***	-0.0133***	-0.0135***	-0.0149***	-0.0164***	-0.0157***	-0.0158***	-0.0142***	-0.0143***
	(0.0015)	(0.0013)	(0.0011)	(0.0010)	(0.0009)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Constant	-0.1204***	-0.0991***	-0.0899***	-0.0819***	-0.0723***	-0.0648***	-0.0576***	-0.0566***	-0.0526***	-0.0504***	-0.0466***	-0.0432***
	(0.0036)	(0.0031)	(0.0029)	(0.0027)	(0.0026)	(0.0024)	(0.0023)	(0.0023)	(0.0023)	(0.0022)	(0.0022)	(0.0022)
Observations	404,304	447,220	490,566	535,451	578,854	622,130	666,563	708,237	744,778	755,349	773,620	806,223
Number of households	95	97	97	98	99	100	100	100	100	100	100	100
Dates	10/30-11/6	11/6-13	11/13-20	11/20-27	11/27-12/4	12/4-11	12/11-18	12/18-25	12/25-1/1	1/1-8	1/8-15	1/15-22

Effects across time cont'd

	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
VARIABLES	13 wks	14 wks	15 wks	16 wks	17 wks	18 wks	19 wks	20 wks	21 wks	22 wks	23 wks	24 wks
Post treat*Avg	-0.0090***	-0.0090***	-0.0081***	-0.0086***	-0.0081***	-0.0080***	-0.0080***	-0.0079***	-0.0078***	-0.0080***	-0.0078***	-0.0075***
	(0.0010)	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Post treat*Eff	0.0025***	0.0027***	0.0029***	0.0025***	0.0027***	0.0030***	0.0030***	0.0032***	0.0037***	0.0038***	0.0039***	0.0039***
	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Post treat	-0.0088***	-0.0089***	-0.0097***	-0.0097***	-0.0099***	-0.0103***	-0.0133***	-0.0154***	-0.0170***	-0.0179***	-0.0187***	-0.0187***
	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0009)	(0.0008)	(0.0007)	(0.0007)	(0.0006)	(0.0006)
Avg ref group	0.0040***	0.0040***	0.0041***	0.0044***	0.0041***	0.0041***	0.0043***	0.0045***	0.0046***	0.0046***	0.0045***	0.0045***
	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)
Eff ref grp	-0.0033***	-0.0027***	-0.0022***	-0.0022***	-0.0021***	-0.0022***	-0.0020***	-0.0018***	-0.0017**	-0.0015**	-0.0014**	-0.0013*
	(0.0008)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)
Children	0.0014***	0.0010***	0.0015***	0.0013***	0.0015***	0.0011***	0.0010***	0.0007***	0.0006***	0.0005**	0.0003	0.0002
	(0.0003)	(0.0003)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Square feet	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Floor	0.0143***	0.0134***	0.0127***	0.0122***	0.0116***	0.0112***	0.0107***	0.0104***	0.0101***	0.0098***	0.0096***	0.0094***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Env org member	-0.0145***	-0.0149***	-0.0144***	-0.0148***	-0.0150***	-0.0152***	-0.0155***	-0.0153***	-0.0154***	-0.0153***	-0.0153***	-0.0151***
	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Constant	-0.0395***	-0.0346***	-0.0276***	-0.0242***	-0.0216***	-0.0193***	-0.0142***	-0.0112***	-0.0087***	-0.0069***	-0.0049***	-0.0040***
	(0.0021)	(0.0020)	(0.0020)	(0.0019)	(0.0019)	(0.0019)	(0.0017)	(0.0016)	(0.0015)	(0.0014)	(0.0014)	(0.0014)
Observations	844,317	886,570	928,053	969,412	1,011,871	1,054,156	1,098,131	1,141,444	1,183,943	1,225,600	1,266,269	1,297,780
Number of households	100	100	100	100	100	100	101	101	101	101	101	101
Dates	1/22-29	1/29-2/5	2/5-12	2/12-19	2/19-26	2/26-3/5	3/5-12	3/12-19	3/19-26	3/26-4/2	4/2-9	4/9-16

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Not reported: Controls for temperature, daylight savings, day of week, week in study, hour in day.

Table A1-6. Regressions including variable coding for participation in previous year's study

	(1)	(2)	(3)	(4)
VARIABLES	total_energy_15min	total_energy_15min	total_energy_15min	total_energy_15min
Post_Treatment_Ave_inFinHealth			0.0306***	
			(0.0011)	
Post_Treatment_Cons_inFinHealth			0.0089***	
			(0.0011)	
Post_Treatment_Ave_inFinHealthF				0.0121***
				(0.0010)
Post_Treatment_Ave_inFinHealthH				0.0287***
				(0.0012)
Post_Treatment_Cons_inFinHealthF				-0.0104***
				(0.0010)
Post_Treatment_Cons_inFinHealthH				0.0043***
				(0.0009)
Post_Treatment*Average	-0.0091***	-0.0063***	-0.0318***	-0.0126***
	(0.0008)	(0.0008)	(0.0012)	(0.0008)
Post_Treatment*Efficient	0.0025***	0.0044***	-0.0058***	0.0053***
	(0.0008)	(0.0007)	(0.0012)	(0.0008)
Post_treatment	-0.0090***	-0.0104***	-0.0086***	-0.0101***
	(0.0013)	(0.0013)	(0.0013)	(0.0013)
Average reference group	0.0022***	-0.0060***	-0.0003	-0.0081***
	(0.0007)	(0.0007)	(0.0007)	(0.0008)
Efficient reference group	-0.0034***	-0.0119***	-0.0058***	-0.0134***
	(0.0007)	(0.0007)	(0.0007)	(0.0007)
inFinHealth	-0.0121***		-0.0279***	
	(0.0005)		(0.0009)	
inFinHealthFin		-0.0159***		-0.0181***
		(0.0004)		(0.0006)
inFinHealthHealth		-0.0159***		-0.0212***
		(0.0004)		(0.0006)
Children	-0.0000	-0.0005**	0.0007***	-0.0008***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
SquFt	0.0001***	0.0001***	0.0001***	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Floor	0.0084***	0.0096***	0.0085***	0.0095***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Env org member	-0.0178***	-0.0163***	-0.0177***	-0.0153***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Constant	0.0110***	0.0157***	0.0315***	0.0148***
	(0.0021)	(0.0020)	(0.0022)	(0.0020)
Observations	1,297,780	1,297,780	1,297,780	1,297,780
Number of l_id	101	101	101	101

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Not reported: Controls for temperature, daylight savings, day of week, week in study, hour in day.

N1. Generalizability of the Sample

In terms of generalizability, our sample consists of Los Angeles Department of Water and Power (LADWP) customers who pay their electricity bills and who are generally representative of California multi-family renter populations, both in terms of the housing characteristics and many common demographic categories (per capita electricity usage, age, income, household composition, etc.), but with higher education levels. Table A1-1 in Appendix presents a comparison of the per capita electricity consumption of our sample. We show that our sample is representative of a general LADWP electric utility consumer, slightly below the general California consumer, and below the national average due to the milder climate in California. The multi-unit housing characteristic of our sample is also generally representative of a broader population. For example, 42.1% of housing units in Los Angeles County and 30.9% of housing units in California are multi-unit structures, making the multi-unit setting a meaningful one to study (U.S. Census Bureau, 2014). More generally, we note that there are 28.1 million multi-family housing units in the United States (Residential Energy Consumption Survey 2013, 2009 data) and 24.3 million of these housing units are renter occupied. These multi-family renter populations do represent a large addressable population. We also note that 90% of all multi-family housing units in the United States are 1, 2 and 3 bedroom units (RECS 2009), with the most common type being 2-bedrooms (there are 12.7 million 2-bedroom units in the United States). According to the American Community Survey, the plurality of households (30.3%) has 2 bedrooms, the mode in our sample as well. Renters occupy 52.7% of American households (American Community Survey Office, 2013). In those rented households, the average number of occupants was 2.84 people, which falls very close to the average occupancy of 2.42 people in our sample. In our sample, all multi-family apartments are 1,2 and 3 bedroom units, with 2-bedrooms being the most common type (N=84 households, 85% of all units in the study). In terms of sample demographics, the median age in our sample of participants (heads of household) is 31 (ranges from 22 to 47) and slightly lower to the median age in California (35.2), and in the U.S. (37.2) (U.S. Census 2010). We note that persons aged 18 to 44, of which our sample participants are representative of, make up 38.7% of the entire population in California (14.4M people), and 36.5% of the U.S. population (112.8M people) based on Census data. In terms of educational attainment status, our participants at University Village are more highly educated than the general U.S. population, having all received a bachelor's degree or higher. We note however, that this is still a population of interest. Persons with a bachelor's degree or higher (age 25+) represent 29.5% of the population in Los Angeles county and 30.5% of the population in California and 31.7% of the population in the U.S. as a whole (U.S. Census 2010).

N2. Robustness Tests

First, in randomized trials, intention to treat often differs from actual treatment. That is to say, some participants might not receive the treatment. In our experiment, we were able to observe whether our participants were actually treated by accessing information on how often they opened the email messages and visited the dashboard. We sent out 24 emails (1 per week) to each participant. On average, these emails were opened 13 times, a little bit more than 50% of the time. Interestingly, we did not find significant differences between the average, efficient groups (Average 13.037; Efficient 12.727). In addition, our participants could access the dashboard any time; the results from Google Analytics indicate that participants visited the dashboard an average of 11 times over the course of the study. Again, we did not find significant differences by group (Average 11.78, Efficient 11.09). In our sample, 5 participants did not open their email nor visited the website. We ran two analyses to test for the inclusion of these participants presented in Table A1-2 in the Appendix. First, we included a dummy that represented whether any email was opened (Model 1). This variable proved to be negative and significant, indicating that participants who saw the treatment tended to consume less than those who didn't. The treatment effects (average and efficient) were consistent with our main results in Table 1-2. Second, we ran the analysis without those participants (Model 2). Again, the results are similar to our main results presented in Table 1-2. We preferred to keep the main results, which represent an intention to treat analysis, since they avoid various misleading artifacts that can arise in intervention research such as non-random attrition of participants from the study or crossover. Furthermore, it is still possible that these participants did see the information by using the preview function of their email and without actually clicking on the email.

Second, since some of our participants were involved in a previous energy use feedback experiment, we tested whether how the inclusion of these participants impacted our results. We found that our treatment effects are robust to controlling for inclusion in the prior study. Additionally, households involved in the prior study tended to consume less overall.[7]

Third, while the effects of valence and distance provide an initial first step to understanding why the average reference group treatment was more effective than the efficient reference group treatment in the present study, the coefficients reported cannot be interpreted too stringently because these analyses are subject to a potential identification problem (Manski, 2003). Specifically, the valence of feedback can change for each household from week to week, and a household's performance in a given week can affect the feedback they receive the following week (which in turn can affect their performance and so on).[8] Because of this potential issue, we also employed a population average approach using the generalized estimating equation (GEE), which is robust to unknown correlation between outcomes (Hubbard et al., 2010; Davidian, 2007). For this test, we had to aggregate up to hourly kWh observations, as the 15-minute data exceeded Stata/MP matrix size limits. The trend remains similar to our prior results (see Table 1-4); receiving favorable feedback was associated with a reduction in consumption ($z = -4.47$, $p < .001$, CI_{95}: -.0013, -.0005) and receiving unfavorable feedback was associated with an increase in consumption ($z = 2.99$, $p = .003$, CI_{95}: .0002, .0008). We also found that increasingly favorable feedback (i.e., increasing distance) is associated with a significant decrease in consumption ($z = -2.26$, $p = .024$, CI_{95}: -.0022, -.0002) and increasingly unfavorable feedback is associated with a significant increase in consumption ($Z = 8.45$, $p < .001$, CI_{95}: .0008, .0013). The fact that we had consistent findings using both a GLS and population average approach lends additional confidence to the robustness of our results.

N3. Other Moderators

Impact of Environmental Factors

We were also interested to see whether our main reference group results might vary by environmental factors such as weather, time of day (daytime vs. nighttime), and time of week (weekdays vs. weekends); these results are reported in Table A1-4. We first compared observations in which the weather was warm or hot (>65 degrees Fahrenheit; Model 1) to

[7] Results available upon request from the authors.

[8] Stated another way, the expected value of the product of feedback in one week and the error in the next week is a non-zero value because the expected value of the product of error terms in the same two weeks is non-zero.

observations in which the weather was colder (<65 degree Fahrenheit; Model 2). We chose this cutoff because 65 degrees distinguishes between heating degree days and cooling degree days. Our main finding for average reference group households was replicated during colder temperatures only, while our main finding for efficient reference group households was replicated during warmer temperatures only. Specifically, when the weather was warmer, the average reference group treatment did not result in a significant consumption change, β= -.00086, z = -.43, p = .669, while the efficient reference group treatment resulted in a significant increase, β= .016, z = 8.20, p < .001. Conversely, when the weather was colder, the average reference group treatment resulted in a significant decrease in consumption, β= -.0091, z = -7.43, p < .001, while the efficient reference group treatment did not result in a significant change, β= .00056, z = .48, p = .634. This suggests that households in the average reference group tended to enact the bulk of their conservation behavior during colder temperatures, while households in the efficient reference group increased consumption during warmer temperatures. We note that an analysis of the temperatures during the treatment period indicated that temperatures tended to be on the colder side (for California) during the treatment with 64 days falling below 50 degrees.

We next compared daytime observations (Model 3) to nighttime observations (Model 4). We chose to define daytime as 7 am-7 pm based on the fact that consumption tended to increase around 7 pm, suggesting that is when residents tended to return home. Our main result was replicated during the day only; that is, during the day, the average reference group treatment resulted in decreased consumption, β= -.0145, z = -12.57, p < .001, while the efficient reference group treatment resulted in increased consumption, β= .0080, z = 7.97, p < .001. However, these results did not emerge at night; there was no significant difference in consumption among households in the average reference group treatment, β= -.00022, z = -.23, p = .820, or among households in the efficient reference group treatment, β= -.00084, z = -.82, p = .414. This suggests that the majority of behavior change occurred during the day. Because residents were less likely to be home during this interval, it is possible that their consumption changes reflect modifications in default settings, such as setting the thermostat higher or lower upon leaving the house. Finally, we compared observations during weekdays to observations on weekends to see whether consumption changes depended on time of the week. Our main result was replicated during both weekdays and weekends. On weekdays, the average reference group treatment resulted in a significant decrease in consumption, β= -.0063, z = -7.09, p < .001, while the efficient reference group treatment resulted in a significant increase in consumption, β= .0040, z = 4.69, p < .001. Similarly, on weekends, the average reference group treatment resulted in a significant decrease in consumption, β= -.0092, z = -5.66, p < .001, while the efficient reference group treatment resulted in a significant increase in consumption, β= .0031, z = 2.02, p = .044.

Effects Across Time

Finally, we wanted to see how the effects of treatment varied over time. Because feedback was emailed weekly, we chose to examine effects by cumulative week (i.e., first week of treatment, first two weeks of treatment, up to the total 24 weeks of treatment), presented in Table A1-5. Cumulative energy savings are usually the performance metric chosen to measure the effectiveness of behavioral interventions. Interestingly, effects did not emerge immediately; in the first week, those in the average reference group treatment, β= .0033, z = 1.34, p = .181, did not show significant changes, although those in the efficient reference group treatment, β=

.0072, $z = 3.15$, $p = .002$, did. It is possible that in this first week, participants were still unsure which behavioral changes would result in significant difference in energy consumption. However, we see that the main finding emerges by the second week of treatment; the average reference group treatment resulted in a significant decrease in consumption, $\beta = -.0038$, $z = -2.06$, $p = .039$, while the efficient reference group treatment resulted in a significant increase in consumption, $\beta = .0090$, $z = 5.18$, $p < .001$. This effect remains consistent throughout the rest of the treatment period, except for treatment weeks 9 and 10, in which the efficient reference group treatment does not result in significant differences. However, this atypical result may be due to higher than usual missing data during that interval; subnetwork configuration changes beyond our control triggered failures in the monitoring systems, which took time to resolve and repair.

CHAPTER 2

Algorithm Aversion and the Aversion to Counter-Normative Decision Procedures

Jonathan E. Bogard

Anderson School of Management, University of California, Los Angeles

Suzanne B. Shu

SC Johnson College of Business, Cornell University

ABSTRACT—According to Norm Theory, decisions that turn out badly result in greater levels of regret if they stemmed from a non-normative decision. Algorithm Aversion (AA) holds that people penalize errors made by algorithms more than errors made by humans. Often, though, studies of AA have explored contexts in which utilizing algorithmic decision-making is unconventional, confounding these two psychological forces. Across five studies, we show that much of what appears as AA can instead be explained by an aversion to counter-normative decision procedures. We find that algorithms are excessively penalized to the extent that using an algorithm to make a forecast is uncommon for that particular domain. In fact, when algorithms are the common decision procedure, we reverse AA and observe a *preference* for algorithms. Using these insights, we can decompose apparent AA into a combination of an aversion to unconventional decision procedures and a residual aversion to algorithms themselves. Overwhelmingly, the larger effect seems to be an aversion to uncommon decision procedures. This investigation offers insight into the mechanism driving AA, explains why people are sometimes averse to algorithms and other times favor them, and suggests a strategy for increasing the utilization of algorithms to improve wellbeing.

Keywords: algorithm aversion, norms, norm theory

1. Introduction and Literature Review

Imagine two people, Anderson and Burns, each separately running late in a cab ride to the airport, each following their respective usual routes. Suppose Anderson decides to deviate from the usual route upon the recommendation of the cab driver, and suppose Burns also deviates from the usual route but upon the recommendation of Google Maps. Each misses their flight. How likely is it that Anderson vows to never follow the advice of a cab driver? How does this compare to the chances that Burns never trusts Google Maps again? Here we show evidence that, under certain circumstances, contrary to recent findings, people are more likely to penalize an erring human than an erring algorithmic forecaster.

A vast and growing literature has documented that, in many cases, people seem to prefer human forecasters over algorithmic ones (Dawes, 1979; Dawes et al., 1989; Dietvorst et al., 2015). This Algorithm Aversion ("AA") has been observed even in cases when the algorithmic forecaster obviously and considerably outperforms the human forecaster (Dietvorst & Bharti, 2020). People often seem to unduly penalize imperfect algorithms more than human forecasters who make the same or even larger mistakes. This bias, to the extent that it exists, proves an obstacle to well-being: Across a variety of domains important to people's lives, algorithmic forecasters have been shown to outperform human decision makers (Dawes et al., 1989; Grove et al., 2000), improving well-being whenever implemented (Gates et al., 2002; Kleinberg et al., 2018). In this way, an AA bias makes people worse off. It is thus important to understand the source of an aversion to algorithms to determine how such an aversion might be overcome in order to increase utilization and thus well-being.

How, though, to reconcile the claims of AA with the apparent ever-growing spread and acceptance of algorithms (Rainie & Anderson, 2017)? Indeed, it is the case that jobs, forecasts, and decisions that were once held by humans are increasingly being displaced by algorithms (Demetis & Lee, 2018). From proofreading to navigation, from cancer screening to romantic matchmaking, tasks once exclusively executed by humans are increasingly being performed by algorithms. If people were unequivocally averse to algorithms, we might expect that algorithms could not gain such prominence. But, as algorithms proliferate and embed further in quotidian decision-making, some have even observed a *preference* for algorithms to human forecasters (Logg et al., 2019). What explains this apparent puzzle, both an aversion to algorithms and widespread usage of algorithms?

One clue, we suggest, can be found in the descriptive social norm favoring one forecasting method or another for a given decision domain. In this paper, we argue that the extent to which people prefer humans to algorithmic decision makers depends on what the normative decision procedure is. When the descriptive social norm favors humans making a given decision, we expect to observe an aversion to algorithms. However, when the norm instead favors algorithmic decisions—as has come to be the case in the domain of navigation, among many others—we expect to see this aversion shrink or even reverse. This is why we expect that, in the opening thought experiment, people will be quicker to abandon a cab driver than Google Maps: Navigation by GPS has become so normalized that for many people it is strange to imagine *not* using an algorithmic decision procedure to navigate.

People's decisions often conform to the prevailing social norm (Cialdini et al., 1991; Cialdini & Goldstein, 2004; Cialdini & Trost, 1998). Norm Theory suggests that this might occur because people have stronger negative reactions when a bad outcome results from a counter-normative decision than a norm-adhering decision (Feldman & Albarracín, 2017; Kahneman & Miller, 1986). For instance, if two people both pick up hitchhikers and both are robbed, most people expect that the person who never picks up hitchhikers will feel greater regret than someone who routinely does. Across multiple domains, people expect greater regret to follow from counter-normative decisions that turn out badly than norm-adhering decisions with the same outcome. Explaining this, Kahneman and Miller offer that counterfactual outcomes are more readily available—thus easier to vividly imagine—for norm-deviating choices than norm-adhering ones. The ease of this "cognitive editing" process, imagining how the decision could have turned out differently, facilitates greater expected regret and a more intense negative affective response. This pattern is in line with Kahneman and Tversky (1982), who find that abnormal causes give rise to stronger affective responses due to the greater availability of a counterfactual. Relatedly, Miller and McFarland (1986) find that juries will award larger compensations to victims in an unusual circumstance rather than a normal one. Zeelenberg and colleagues find direct evidence supporting the claim that expected regret stems from behavior-focused counterfactual thinking (Zeelenberg et al., 1998). Taken together this research suggests that, should a decision turn out badly, counter-normative decisions are likely to engender greater anticipated regret than norm-adhering decisions. This implies that algorithms will be expected to elicit greater anticipated regret when they are seen as counter-normative.

Researchers have observed that people often make uncertain decisions using an expected-regret-minimization strategy (Simonson, 1992; Larrick & Boles, 1995; Ritov, 1996; Zeelenberg, 1999). People's choices under uncertainty are especially likely to minimize anticipated post-outcome regret when feedback about the event's outcome is expected (Zeelenberg & Pieters, 2004). Combining this tendency with the findings of Norm Theory, it follows that norms affect choice via anticipated regret. Counter-normative options are likely to engender greater anticipated regret, and this greater anticipated regret is likely to impel people toward choosing the normative option. In a test of this idea, Bar-Eli and colleagues (2007) find that a prevailing norm among elite goalkeepers in favor of diving to either side when defending penalty kicks leads to sub-optimal decisions for the sake of regret-minimization. In line with this, we expect people to be averse to algorithms to the extent that the norm favors a human decision maker, often irrespective of whether or not this is the optimal strategy.

2. Current Research

While various attempts have been made to mitigate an apparent psychological aversion to algorithms (Dietvorst et al., 2018; Schmidt et al., 2020), relatively little attention has been paid to understanding the source of the aversion in the first place. The purpose of the present investigation is to better understand the mechanism by which algorithm aversion occurs, and also to help explain when people are likely to be averse, indifferent, or attracted to algorithmic forecasters.

Because counter-normative decision procedures cause stronger negative reactions and greater levels of anticipated regret than norm-adhering ones, and because people often make decisions in ways that minimize anticipated regret, we expect that people will avoid algorithms whenever use of algorithms is against the norm. Thus, we hypothesize the following:

H_1: Anticipated regret for algorithmic (versus human) forecasters will be higher when use of an algorithm is inconsistent (versus consistent) with the descriptive social norm.

H_2: Willingness to use imperfect algorithmic (versus human) forecasters will increase when use of an algorithm is consistent (versus inconsistent) with the descriptive social norm.

H_3: Anticipated regret will mediate the effect of norm-adhering (versus norm-deviating) decisions on subsequent selection of a forecasting procedure.

We test these hypotheses in the domains of health forecasts, navigation, gambling, and forecasts of professional performance, using dependent measures of anticipated regret, hypothetical personal decisions, and incentive-compatible choice. By measuring or manipulating the prevailing norm for which forecasting procedure to use, we are able to replicate traditional algorithm aversion, minimize it, and even reverse it. Throughout, we use “norm” to mean “descriptive social norm” (i.e., the most common choice) rather than referring to the injunctive norm (i.e., what is recommended) or what is rational or best for a person (i.e., what is meant when the optimal option in a choice set is described as the “normative” decision). For all studies in this paper, all procedures, analyses, and sample decisions were pre-registered and are available on Research Box.[9] This investigation offers insight into the mechanism driving an aversion to algorithms, explains why people are sometimes averse to algorithms and other times favor them, and suggests a strategy for increasing people’s usage of algorithms to improve wellbeing.

3. Experiment 1: Anticipated Regret in Bloodwork Analysis

Norm Theory research has shown that people expect greater regret from a decision that turns out badly if an unconventional option was chosen. If using an algorithm is seen to be against the norm then a pattern similar to AA would appear, but it could be driven to some extent by an aversion to unconventional decision procedures and not the use of algorithms *per se*. As such, in Experiment 1, we measured the differences in expected regret from using an algorithmic or a human decision procedure while experimentally manipulating whether this procedure was consistent or inconsistent with the norm.

[9] https://researchbox.org/580&PEER_REVIEW_passcode=IZGCVR

3.1 Method

3.1.1. Participants. In seeking to power our results at a level similar to Dietvorst et al. (2015), we recruited 1200 American residents through Amazon's Mechanical Turk (MTurk) platform who had MTurk user ratings of at least 95%. After removing from analysis any person who failed any of the four pre-registered attention checks, we were left with 1168 participants (50% female, $M_{age} = 38.8$, $SD_{age} = 12.3$).

3.1.2. Procedure and Measures. Participants were all introduced to Smith, a person who was not feeling well for several weeks so went to the doctor. The doctor worried that Smith might have a rare blood condition and drew a blood sample to send to a local hematology lab. Participants were then randomized into one of four experimental conditions, corresponding to our 2 (Smith's choice: human or algorithm to perform the analysis) x 2 (Norm: the decision procedure that Smith chose was norm-consistent or norm-inconsistent) between-subjects design. Subjects read:

> For this particular condition, the bloodwork is nearly *always* analyzed by a computer algorithm [doctor], and the results are 99% accurate, but Smith *could* choose to have the results analyzed by a doctor [computer algorithm] instead.
> In this case, Smith decides to have the computer algorithm [doctor] analyze the blood results.

In all cases, participants were told that the test concluded that Smith does not have the blood condition and so Smith was sent home. Participants were then informed that Smith fell ill weeks later, that the analysis of the bloodwork turned out to be wrong, and that Smith does in fact have the blood disorder. Thus, all participants were given information about the norm, were told about Smith's decision to use a human or algorithm to analyze the blood-work—which was either consistent or inconsistent with the norm—then told that the chosen procedure had erred.

Similar to research in Norm Theory, participants were then asked how much regret they expected Smith to feel about the decision to choose a human or an algorithmic forecaster (0 = no regret at all, 3 = a moderate amount of regret, 6 = a lot of regret). We were interested in differences in the amount of expected regret after observing an imperfect human versus algorithmic forecast, and tested whether this expected regret depended on whether the decision was consistent or inconsistent with the norm for this particular judgment. After the measure of expected regret, participants answered the attention check questions and indicated their age and sex.

3.2 Results

3.2.1. Main Analysis. We were interested in whether the amount of expected regret resulting from an imperfect human versus algorithmic forecast depends on whether or not the observed forecast was consistent with the norm. To test this, we estimated the following model:

$$regret \sim \beta_1 * observed\_algorithm + \beta_2 * norm\_inconsistent$$

Here, *regret* is a continuous measure, *observed_algorithm* corresponds with the experimental condition of whether participants read that Smith opted to have the doctor or the algorithm analyze the bloodwork, and *norm_inconsistent* is a binary variable corresponding to whether Smith's decision was said to be consistent with the norm (i.e., when the norm was human analysis and Smith chose the human, or when the norm was algorithmic analysis and Smith chose the algorithm) or inconsistent with the norm (i.e., when the norm was human analysis and Smith chose the algorithm, or when the norm was algorithmic analysis and Smith chose the human). As is the case for all experiments presented in the current investigation, higher values on the dependent measure represent a greater penalty (here, via higher regret) for the given decision procedure.

We first examined whether there was a significant main effect of norm inconsistency on expected regret. In line with our prediction in H_1, we find a highly significant impact of norm inconsistency ($\beta_{norm\_inconsistent} = 1.14$, $t = 12.62$, $p < .001$). Substantively, this means that, on a

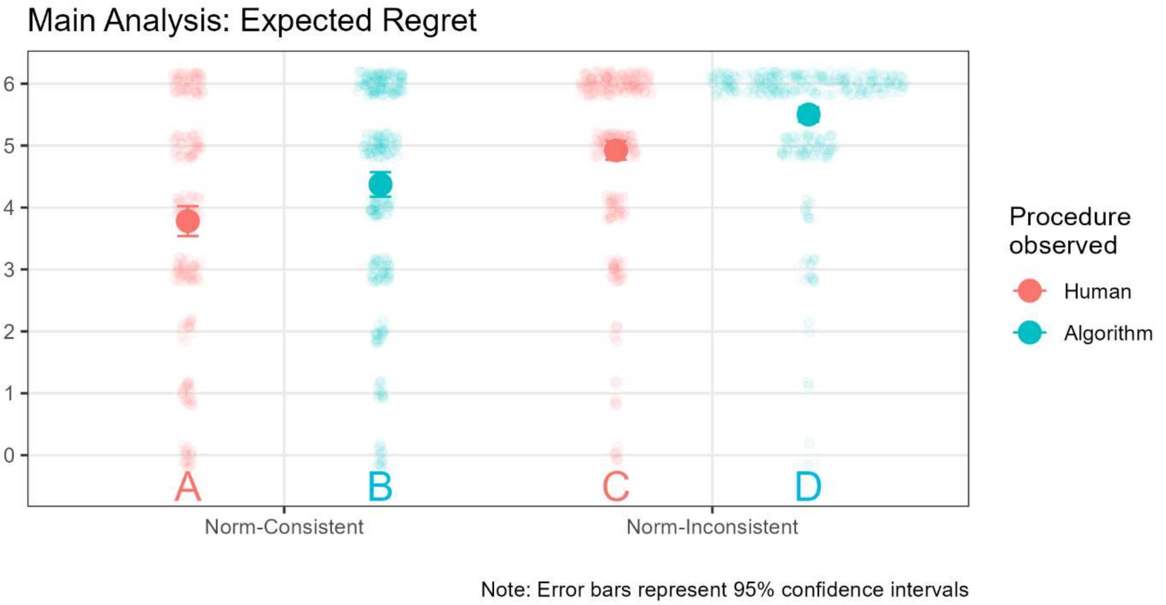


Figure 2-1. *Expected regret resulting from a norm-consistent versus norm-inconsistent selection of forecasting method, whether algorithmic or human.*

Seven-point scale of regret, the additional regret expected from a counter-normative decision procedure is more than a full point higher on average *whether or not* this was from a human or algorithmic procedure. Further, we take β_1 from the model specification above, the main effect of using an algorithm, as a cleaner test of AA while controlling for the norm. We find that there does appear to be a residual additional penalty imposed on imperfect algorithms while controlling for the norm ($\beta_{observed\_algorithm} = 0.58$, $t = 6.51$, $p < .001$), but that the effect of norms is nearly double the effect of algorithm aversion. A linear hypothesis test confirms that this difference is highly significant ($F(1)=20.43$, $p< .001$). This is a first step toward understanding how much of people's apparent aversion to algorithmic decision procedures is an aversion to algorithms *per se* and how much is due simply to the fact that these decision procedures are often counter-normative.

3.2.2. Additional Analyses. We were also interested in additional comparisons made possible by our experimental design to contextualize these effects. For the additional analyses of Experiment 1, we estimated a model similar to the one described above but using slightly different planned contrast-coding of variables and thus a slightly different model specification:

$$regret \sim \beta_1 * observed\_algorithm + \beta_2 * norm\_human + \beta_3 * observed\_algorithm * norm\_human$$

Here, we capture *norm_human* to test the effect of the norm being a human (rather than, as before, the decision procedure being norm-consistent). As described below, we additionally estimate an equivalent model as this one, instead using *norm_algorithm* rather than *norm_human*, reversing the dummy-coding of the *norm* variable.

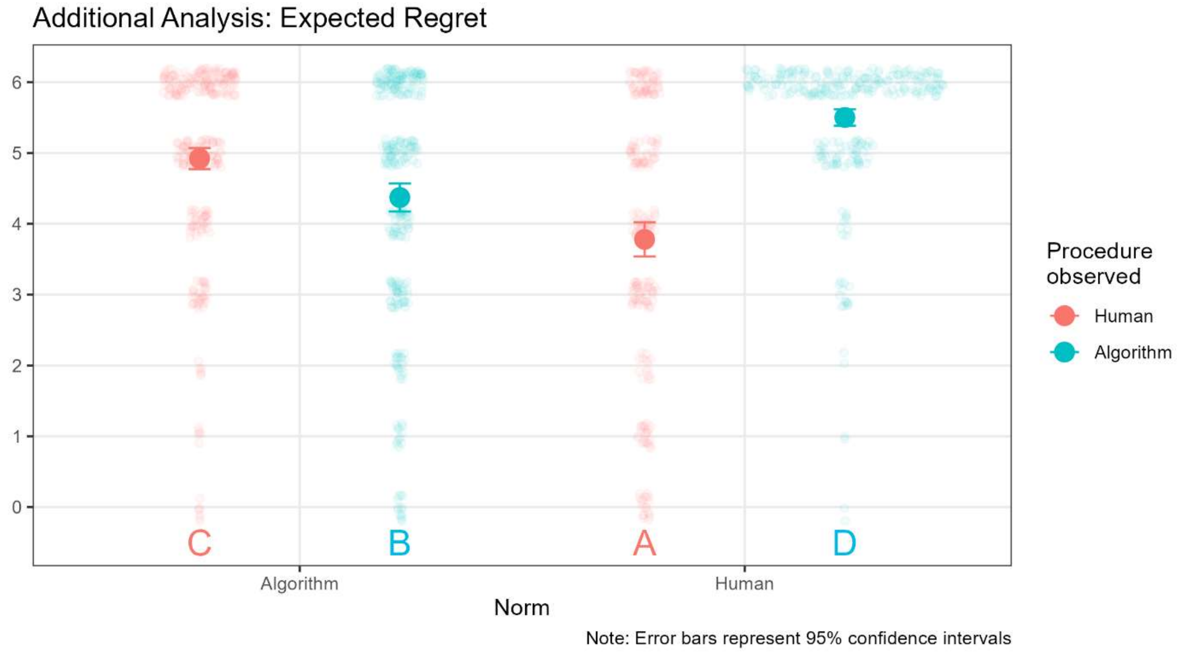


Figure 2-2. *A rearranging of the same data from Figure 2-1, here depicting expected regret from observing a human versus algorithmic decision procedure when the norm is for an algorithm versus human to forecast.*

As is evident when comparing Figure 2-1 and Figure 2-2, all we have done is to rearrange the comparisons by recoding the variables. We have left the lettering of the conditions consistent for easier comparisons. Recoding in this way, however, makes additional comparisons more straightforward to interpret.

3.2.2.1. ***Replicating AA.*** We speculate that much of the AA literature has confounded the use of algorithmic forecasters with the use of non-normative forecasters. That is, for many of the domains in which AA has been observed, there is a strong norm of humans—not algorithms—to make the decision or forecast. This is tantamount, in our study, to Comparing Condition A with Condition D from the Figure 2-2: the difference between observing an imperfect human or algorithm when the norm is for a human to forecast. Indeed, we replicate the typical pattern observed as AA. For decisions in which human forecasters are the norm, we see a considerable aversion to algorithms measured as greater expected regret ($\beta_{\text{observed\_algorithm}} = 1.72$, $t = 12.99$, $p < .001$), replicating the standard AA finding.

3.2.2.2. ***Moderating AA.*** The AA hypothesis holds that imperfect algorithms are penalized more than imperfect human forecasters. While we indeed find this to be true when directly comparing Conditions A and D, we were interested to see whether this effect is moderated by changing the norm for the decision. The variable of interest on the model specification from section 3.2.2 is the interaction term, β_3, representing the extent to which the effect of observing an algorithm on subsequent regret depends on the norm being a human. Graphically, considering Figure 2-2, this comes to comparing the difference-in-differences between Condition C and Condition B (observing a human versus algorithm when the norm is for an *algorithm* to forecast) versus Condition A and Condition D (observing a human versus algorithm when the norm is for a *human* to forecast). Here, we find a considerable cross-over interaction ($\beta_{\text{interaction}}$ = -2.27, t = -12.61, p < .001), consistent with our hypothesis that the extent to which people are averse to algorithms crucially depends on the norm.

3.2.2.3. ***Reversing AA.*** As noted above, when the norm is for a human to forecast (Condition A and Condition D), significantly more regret is expected from choosing an algorithm. However, consider the comparison of Condition C and Condition B in Figure 2-2 above—a comparison of observing a human versus an algorithm when the norm is for an *algorithm* to forecast. If the predictions of AA are to be taken literally, we would expect greater regret associated with a failed algorithm compared to a failed human forecaster. That is, we would expect Condition B to be significantly greater than Condition C. However, we instead hypothesize that greater regret will be associated with the counter-normative decision procedure, the human in this case. This is, in fact, the pattern we observed ($\beta_{\text{observed\_algorithm}}$ = -0.55, t = -4.51, p < .001). In other words, when the norm is for algorithms to forecast, instead of observing an aversion to algorithms, we instead observe a significant *preference* for algorithms.

3.3. Discussion

In Experiment 1, we tested whether the amount of expected regret resulting from a medical determination made by an incorrect human versus an incorrect algorithm depends on whether the norm favors a human or an algorithm. In our first analysis, we find two main effects: higher regret from having chosen an algorithm, and higher regret from having used an option inconsistent with the norm. Importantly, the latter effect of norm inconsistency is about twice as large as the effect of using an algorithm. In further analyses we decompose these effects to specifically look at pairwise comparisons of the effect of algorithm aversion for each type of norm. While replicating prior findings of AA when the norm favors human decision makers, we observe a full reversal when the norm favors algorithmic decision makers such that algorithms are preferred. We therefore conjecture that previous demonstrations of AA were observed largely because they were studied in domains where norms favor human forecasters.

4. Experiment 2

In Experiment 1, we find that the normative decision procedure accounts for a large share of expected regret, with an aversion to algorithms demonstrated to be considerably less important. In fact, the total amount of AA observed is moderated by the normative decision procedure. When the norm is for humans to forecast, AA appears. However, when the norm is for algorithms to forecast, the reverse pattern was observed. While Experiment 1 measured the amount of AA by looking at expected regret—a measure common in the Norm Theory literature—we wondered whether a similar finding would obtain when using a dependent measure more common to the AA literature: subsequent choice of a decision procedure. Here, we measure people's penalizing a given decision procedures by observing their choice to use the same decision procedure or to switch procedures from the one that they observed.

4.1 Method

4.1.1. Participants. After detecting a large effect of the norms manipulation in Experiment 1, in Experiment 2 we reduced our target sample size to 600 participants from MTurk and, after a similar set of four attention checks, were left with 548 participants (52% female, M_{age}= 40.6, SD_{age}= 12.7).

4.1.2. Procedures and Measures. Experiment 2 was nearly an identical replication of Experiment 1 with one critical change: we added an additional measure, participants' desire to choose a human or an algorithmic decision procedure if they were to be in a similar position as the protagonist, Smith. We were chiefly interested in two main questions. First, we wanted to see if a similar pattern of results obtained when using a dependent measure of choice as was observed with the regret-based measure. We wanted to evaluate the extent to which people penalized decision procedures—as measured by a decision to switch from the procedure that they observed—when those procedures deviated from the norm. Second, we wanted to see if expected regret, as measured in Experiment 1, mediated this relationship.

To answer these questions, after mimicking the manipulation from Study 1, we then asked subjects who they would choose for a similar analysis to be done for their own health screening (-3 = definitely choose a doctor, 0 = indifferent between an algorithm and a doctor, +3 = definitely choose an algorithm). For participants who observed the forecast of a human, a more positive number on this scale indicates a stronger desire to switch decision procedures (i.e., to use an algorithm instead of the procedure they observed, a doctor). For participants who observed an algorithm's forecast, we reverse-code this choice measure so that more positive numbers again correspond with a desire to switch from the observed procedure (i.e., a stronger personal preference to use a doctor's forecast instead of an algorithm). Thus, on this constructed measure, more positive numbers always indicate a stronger desire to switch from the forecasting procedure that participants observed in their assigned condition. In the results reported below, we binarize the measure such that positive numbers indicate a desire to switch decision procedures (switch=1) and negative numbers indicate a desire to remain with the observed decision procedure (switch=0). This makes interpretation of the results more straightforward. However, using the continuous measure yields a qualitatively identical finding but with higher power because of its finer resolution (see Online Appendix for full analyses). We interpret higher values on this dependent measure to represent a greater penalty of the given decision procedure.

We had two main predictions about people's decisions to stay or switch from the procedure they observed, depending on whether or not this procedure was consistent with the norm. First, per H_2, we predicted that whether participants' choices demonstrated an aversion to algorithms would depend on whether participants observed a norm-consistent or norm-inconsistent decision procedure that erred. We expected people to switch from algorithms much more when use of algorithms was against the norm. Second, in line with H_3, we predicted that the relationship between observed procedure and the decision to switch procedures would be mediated by expected regret, per Norm Theory, as measured as in Experiment 1.

4.2 Results

First, we were interested in whether norm-consistency affects the decision to switch from the observed forecasting procedure, similar to its effects on regret as in Experiment 1. To test H_2, we estimated the following model:

$$decision\ to\ switch \sim \beta_1 * observed\_algorithm\ +\ \beta_2 * norm\_inconsistent$$

We were most interested in β_2 from the model above, representing whether the decision to penalize the observed procedure depends on norm consistency.

As hypothesized, we find a similar pattern of findings as in Experiment 1, this time using a dependent measure of participants' stated subsequent personal choice.

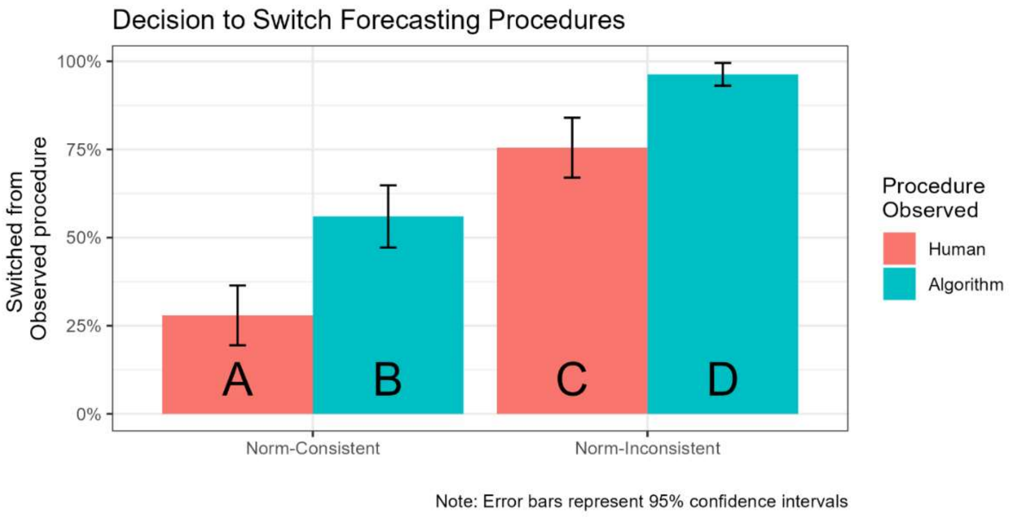


Figure 2-3. *Penalty of algorithmic and human forecasting procedures, as measured by the decision to switch from the observed procedure, depending on whether this was consistent or inconsistent with the norm for that particular decision domain.*

How much participants penalized imperfect forecasting procedures in their subsequent choice, as measured by a decision to switch procedures, depends on whether the observed procedure was norm-consistent or norm-inconsistent ($\beta_{\text{norm\_inconsistent}} = 0.44$, t = 11.70, p < .001). As predicted in H_2, we once again observe a substantial effect of norms on the extent to which AA exists, this time affecting subsequent stated choice preference. As in Study 1, this effect is considerably larger than the independent impact of observing an imperfect algorithm controlling for norm-consistency ($\beta_{\text{observed\_algorithm}} = 0.24$, t = 6.53, p < .001). Once again, the effect of norm-consistency is nearly double the effect of algorithms *per se*.

4.2.1 Additional analyses. Using a model specification similar to that described in Section 3.2.2 but with the choice-based dependent measure of Experiment 2, we once again find a considerable apparent aversion to algorithms when the norm favors human forecasters for that particular decision ($\beta_{\text{observed\_algorithm}} = 0.68$, t = 13.18, p < .001). Furthermore, once again, we reverse the finding—demonstrating a *preference* for algorithms—in the case that the norm favors algorithms to make the determination ($\beta_{\text{observed\_algorithm}} = -0.19$, t = -3.61, p < .001). Last, we replicate the findings from Study 1 when expected regret serves as the dependent measure (see Online Appendix for full discussion).

4.3. Mediation

We hypothesized that expected regret, as measured in Experiment 1, would mediate the relationship between norm consistency and subsequent hypothetical choice to use an algorithm. Indeed, in line with H_3, we do find such a relationship.

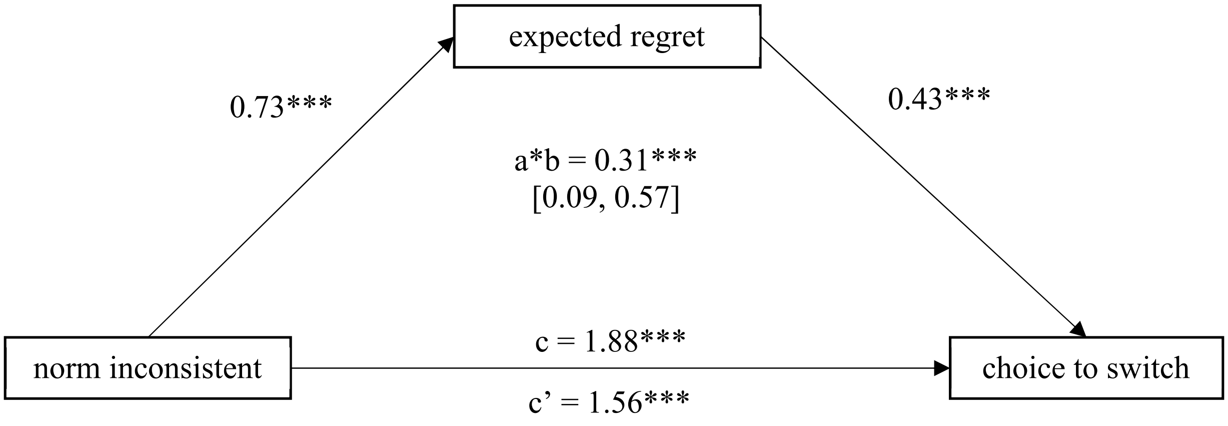


Figure 2-4. *Expected regret mediates the effect of norm consistency on people’s penalty of the forecasting procedure they observe.*

As shown in Figure 2-4, the effect of norm consistency on subsequent decision to switch from the observed procedure is partially mediated by expected regret. Norm inconsistency significantly increased expected regret ($\beta_{\text{inconsistency}}$ = 0.73, t = 4.58, p < .001), which in turn was associated with a greater willingness to switch forecasting procedures (β_{regret} = 0.43, t = 10.84, p < .001). Unstandardized indirect effects were computed for each of 10,000 bootstrapped samples, and the 99.9% confidence interval was computed. A bootstrapped unstandardized indirect effect of 0.31 had a 99.9% confidence interval of [0.10, 0.59], thus expected regret significantly mediates the effect of norm inconsistency on subsequent choice.

4.4. Discussion

Altogether, Experiment 2 replicates the finding from Experiment 1 regarding expected regret and also shows that, similar to regret, participants' stated personal choice to use an algorithm is greatly influenced by the norm. Whether algorithms are penalized more than humans after they each err depends considerably on whether or not the use of an algorithm is normative. We find further that the expected regret resulting from norm-inconsistent decisions mediates this relationship.

5. Experiment 3

Experiments 1 and 2 demonstrate the essential role of norms in shaping people's response to observing imperfect forecasters. Both of these studies were conducted in the consequential and emotionally charged domain of serious health considerations. Further, both scenarios involved a mistake made in diagnosing what might be considered a *loss* of health. The purpose of Experiment 3 was twofold. First, we wanted to test the generalizability of our findings in a new domain. Second, we wanted to test our hypothesis when a failed forecast leads to a foregone gain rather than a realized loss. This, we suggest, is a more conservative test of our hypothesis. Norm Theory holds that the affective response (e.g., regret) to a failed decision is greater when a counter-normative option is selected. Research has shown that foregone gains will be less affect-rich than realized losses (Rottenstreich & Hsee, 2001; Rottenstreich & Shu, 2004). To the extent that, per Norm Theory, our findings depend on there being a more extremely negative emotional response to a failed counter-normative decision procedure, it is a more conservative test of our hypothesis to examine it in the context of a less affect-rich environment: foregone gains. For these reasons, Experiment 3 conceptually replicates Study 1 but in the domain of sports betting.

5.1 Method

Experiment 3 was set up nearly identically to Experiment 1 except the vignette, rather than being about a person who needed bloodwork analyzed, was instead about a person who stood to win $1,000 in a football betting pool if their predictions were more accurate than the other contestants'. In the vignette, the character turned out to have *not* won, and once again the character either used a norm-consistent or norm-inconsistent decision procedure, which was either the algorithm or human forecaster. Note that, unlike in Experiments 1 and 2, the human forecaster was not an outside expert (i.e., the doctor) but the decision-maker themselves. Once again, we predicted that people's aversion to a human or algorithmic forecasting method will depend to a large extent on what the normative forecasting procedure is.

5.1.1. Participants. We recruited a sample of 500 participants from MTurk and, after a similar set of attention checks as in Experiment 1 and Experiment 2, were left with 469 participants (53% female, M_{age} =39.12, SD_{age}=12.05).

5.1.2. Procedures and Measures. Subjects read a vignette about a person, Smith, who paid $10 to join a betting pool to win $1,000 based on the accuracy of predicting the outcome of a season of football games. Subjects were told that the analysis was almost always done by a human [computer algorithm] but that it could instead be forecasted by a computer algorithm [human]. They were then told that Smith chose to go with a human [computer algorithm] forecaster but, at the end of the season, someone else ended up winning the pool.

5.2. Results

As the figures below illustrate, the pattern of results in Study 4 was nearly identical to the pattern observed in Study 1.

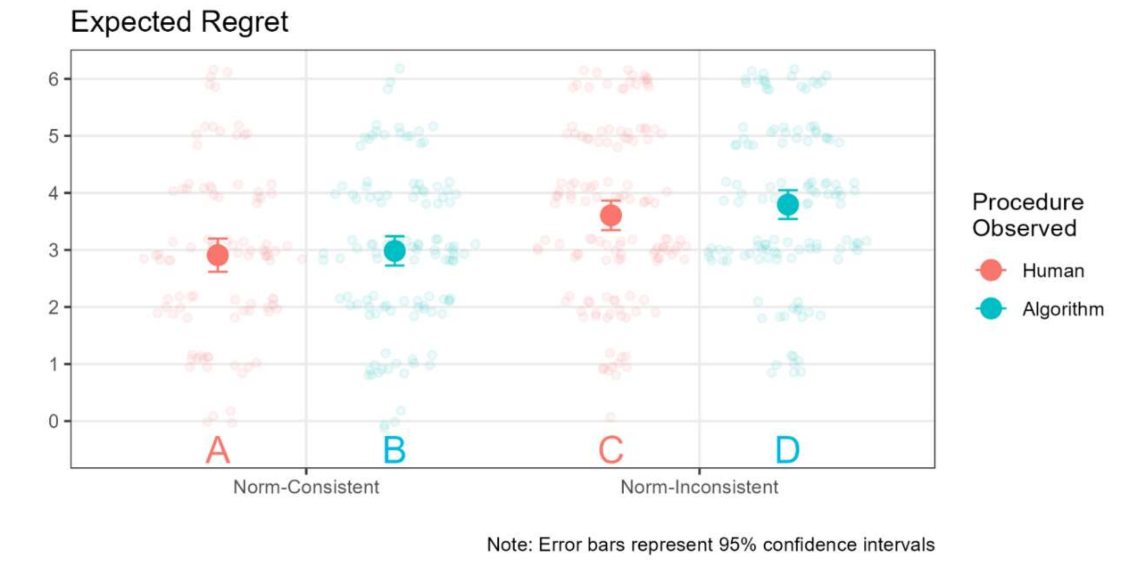


Figure 2-5a. *Expected regret from a norm-consistent versus norm-inconsistent selection of forecasting method, whether algorithmic or human, in the domain of foregone gains.*

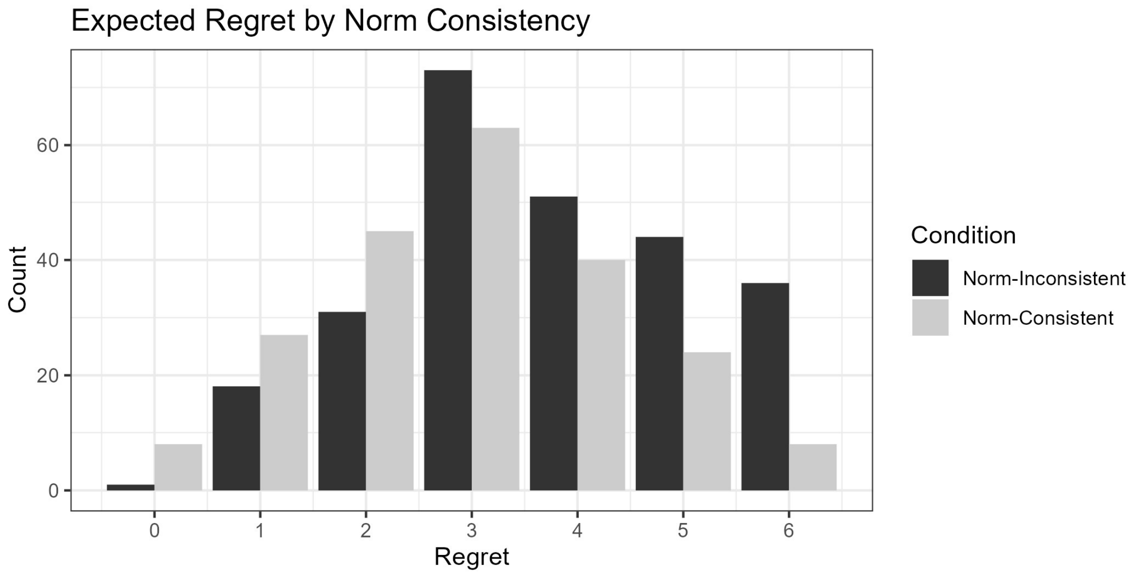


Figure 2-5b. *Distribution of expected regret scores after observing a norm-inconsistent versus norm-consistent forecasting method.*

Most importantly as a test of our main hypothesis, there is a considerable effect of norm-inconsistency on expected regret ($\beta_{\text{norm\_inconsistent}} = 0.76$, $t = 5.65$, $p < .001$), even in the domain of foregone gains. Once again, we additionally observe that the extent to which algorithms are penalized more than humans, as measured by higher regret, depends on whether the norm is for humans or algorithms to make forecasts in this given situation ($\beta_{\text{interaction}}$ = 1.51, t = 5.63, p < .001). Thus, we see a similar pattern of results as in Experiment 1 but in a different, presumably less affect-rich, domain of foregone earnings from gambling.

Note that in Experiment 1, as an additional analysis, we compared Condition A to Condition B as a cleaner test of AA conditional on making a norm-consistent decision. In Experiment 1 we found a residual penalty for imperfect algorithms after accounting for norm consistency. Here, however, we find no such difference, failing to reject the null hypothesis that people expect the same amount of regret when choosing an imperfect algorithmic (versus human) forecaster for norm-consistent decision procedures ($\beta_{\text{choice\_alg}}$ = 0.19, t = 1.04, p = .298, NS). In other words, in this context, we find no evidence of lingering AA after accounting for differences in the norm.

6. Experiment 4

Thus far, for the sake of control, we have experimentally manipulated the stated norm and then looked for an effect of norm consistency. However, there are at least two limitations of this approach. First, people usually infer norms rather than being explicitly told what the norm is, as in Experiments 1 – 3. Being explicitly told about the norm may be driving some of the results observed. Second, it is possible that people are making inferences about the relative accuracy of the human and algorithmic forecaster based on what they were told about the norm. Being told that one procedure (and not another) is the norm may signal that the normative procedure is more accurate. Research has shown that people will often still prefer a human to a higher-performing algorithm (Dietvorst & Bharti, 2020), but we wanted to rule this out as an explanation of our results.

In Experiment 4 we sought to address these concerns and move into yet another domain to further test the generalizability of our findings. In this study, we tested our hypothesis in the domain of navigation by algorithms (e.g., Google Maps) versus human experts (e.g., cab drivers). Interestingly, Dietvorst et al. (2015) use this domain in a thought experiment to advance their conjecture that people penalize imperfect algorithms more than imperfect human forecasters. Whether or not that empirically would have been true at the time of their writing, the contemporary norm clearly seems to favor algorithm-based navigation (Panko, 2018). Beyond all of this, though, we wanted to see if a similar pattern of results would emerge from our measuring—not manipulating—beliefs about the norm. We sought to explore all of these issues in Experiment 4.

6.1 Method

In Experiment 4, participants read a description of two people, each of whom (separately) took a cab to the airport while running late. In both cases, the person agreed to deviate from the original route, but this resulted in them being 20 minutes late and missing their flight. The key difference between the two characters in the vignette is that one of them was offered the alternative route by a cab driver while the other was offered an alternative route by Google Maps. After learning that, separately, each of these alternative routes failed, participants indicated whose forecast they would choose if they were in a similar situation the next time they needed to navigate somewhere while running late (-3 = definitely choose a cab drivers' recommendation, +3 = definitely choose a Google Maps recommendation). Subjects were then asked about whether it is normal to navigate by human recommendation (-3) or Google Maps (+3) in situations like these. We predicted that, after reading a vignette about both a human and an algorithm failing, people's personal decision to use a human- or algorithm-given route recommendation would depend on their perception of how normal it is to navigate by algorithm.

6.1.1. Participants. We recruited a sample of 400 participants from MTurk and, after a similar set of four attention checks as before, we were left with 390 participants (48% female, M_{age}=36.50, SD_{age}=10.18).

6.1.2. Procedures and Measures. Participants read the description of two people, both of whom took a cab to the airport while running late, as described above. Note that there was no between-subjects experimental manipulation. Our key dependent measure was participants' stated preference to use Google Maps or a cab driver to navigate to the airport if they were to be in a similar situation in the near future. We were most interested in whether participants' stated personal preference to use an algorithm for their future decisions correlated with their perception of the norm. To measure this, we asked participants whether they believe the norm in these sorts of situations is to navigate by human or Google Maps recommendations.

6.2. Results

As predicted, after reading about both a human and an algorithm failing, participants were considerably more averse to the Google Maps algorithm to the extent that they saw using Google Maps as uncommon (β_{norm_alg} =0.46, t=4.94, p<.001). As shown in Figure 2-6, individuals who perceived use of Google Maps as the norm were more likely to be willing to use an algorithm for navigation in a future situation.

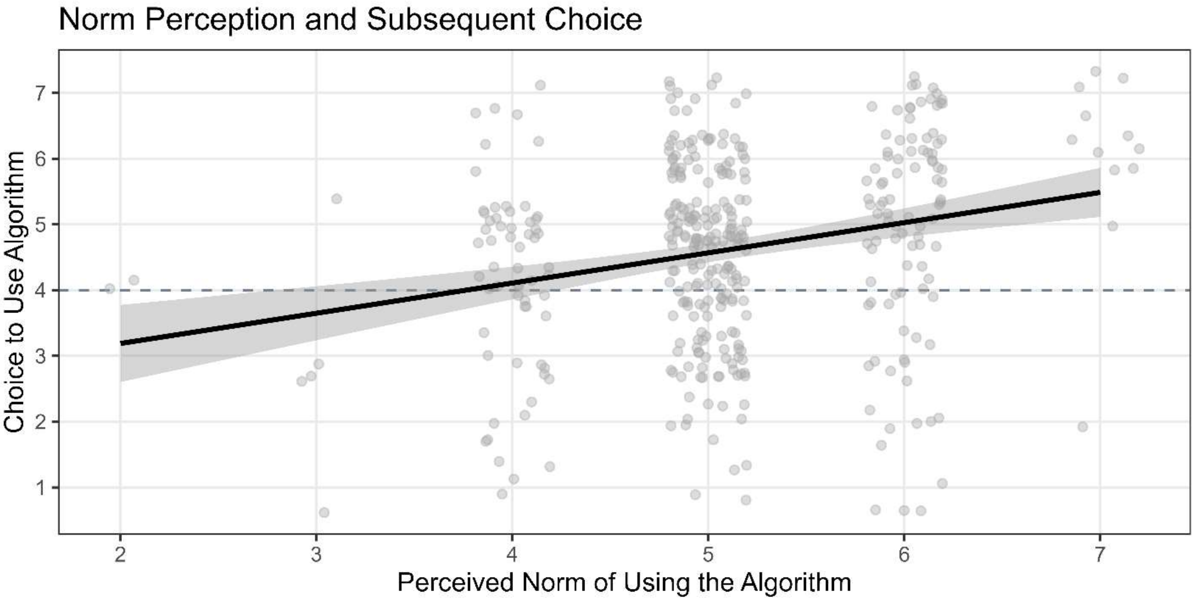


Figure 2-6. *Correlation between decision to use algorithmic navigation and perception that this is the normative procedure.*

Hence, we find further evidence for our hypothesis that AA depends critically on people's perceptions of the norm, and that this is true even when participants' perception of the norm is measured rather than manipulated.

It is noteworthy, in further support of the notion that norms matter to AA, that fully 72.2% of all participants expressed a personal preference for using the algorithm (Google Maps) even after seeing both the algorithm and a human fail at the same task. In a separate pre-test in which subjects were asked the same dependent measure choice question but not presented with a vignette demonstrating an algorithm and a human failing, 71.4% of subjects chose to use the algorithm. Given preferences for the algorithm as high as 71.4% *without* observing failure, prior AA findings suggest that this preference should fall after observing the algorithm's failure. Instead, we find no apparent negative impact on willingness to use algorithms. This gives reason to doubt—at least in familiar domains in which algorithms are overwhelmingly the norm, as with

Google Maps for navigation— that people are unequivocally averse to imperfect algorithms relative to imperfect humans.

7. Experiment 5

Across three separate domains, we find evidence supporting the idea that norms of decision procedures are an essential moderator of AA, sometimes eliminating or even fully reversing the phenomenon. In Experiment 5, we perform a conceptual replication of the Dietvorst et al. (2015) Study 4 involving admissions to an MBA program using an algorithm. Using available materials (see Dietvorst et al. 2015's online supplement), we recreated their experimental design but with two critical changes. First, rather than using a sample of MBA students making decisions about MBA admissions, we decided to replicate using a convenience sample in the domain of professional baseball recruiting. We did this because we speculated that most people do not believe that it is common for an algorithm, rather than human admissions officers, to make admissions and financial aid decisions for MBA programs. Because of the centrality of norms to our theory, we wanted a domain in which we could credibly manipulate the perceived norm. Baseball recruiting seemed like a natural fit since the use of human talent scouts is well known, but the use of algorithm-driven management decisions in baseball has been popularized by the book and subsequent movie *Moneyball*. Thus, Major League Baseball (MLB) performance forecasting offered a domain in which we could credibly manipulate beliefs about the normative decision procedure. The second major departure from Dietvorst et al. (2015) is that in Experiment 5 we provided information about the norm to participants, experimentally manipulating whether the norm was said to be humans (i.e., talent scouts) or algorithms (i.e., sabermetrics forecasts). As in the original, our study was incentive-compatible, with participants' chance at a bonus being tied to the accuracy of their chosen forecasting method.

7.1. Method

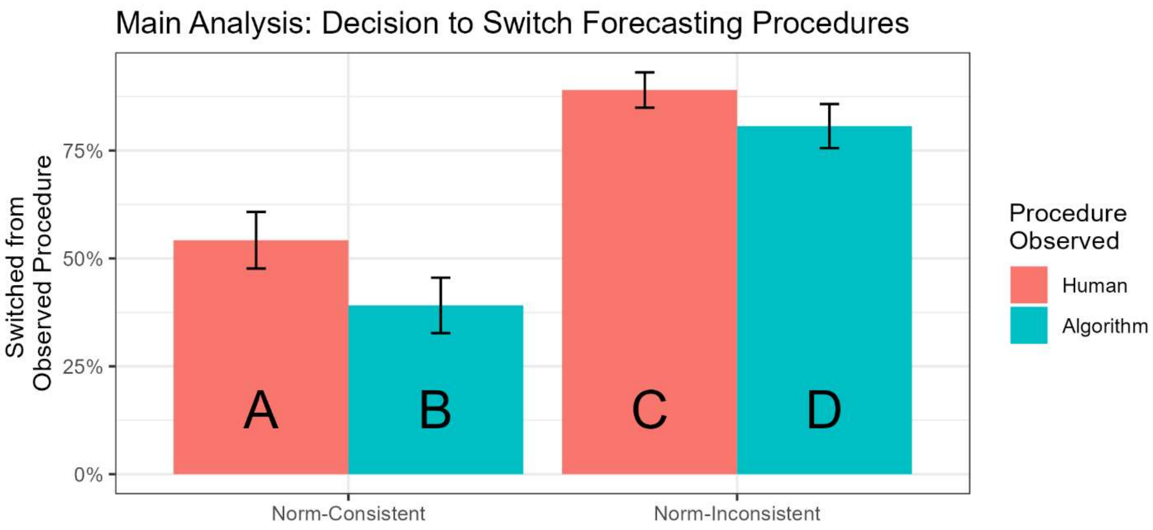
7.1.1. Participants. We recruited a sample of 1000 participants from MTurk and, after a similar set of attention checks, we were left with 911 participants in our 2 (Forecaster: human or algorithm) x 2 (Norm: norm-consistent or norm-inconsistent), between-subjects design (45% female, M_{age}= 39.97, SD_{age}= 12.96). Subjects were paid \$1.25 for participation in our study but could earn additional bonuses according to their forecasting performance.

7.1.2. Procedures and Measures. We structurally replicated the experimental design from Dietvorst et al. (2015): subjects were first told that they would play the role of recruiter for a professional baseball team. Their job was to forecast which players would be most likely to have a successful professional career, defined as an equal weighting of Batting Average, teammate respect (a proxy for nomination to the All Star team), Slugging Percentage, and total runs scored. Each of these terms was defined. Participants were randomized into a human or an algorithm forecaster condition. As in Dietvorst et al. (2015), participants reviewed, one at a time, the past performance of 10 different players summarized on a table of eight attributes. After each player's performance table was shown, participants saw (a) the human or the algorithm's prediction of the player's performance, depending on which condition the participant was randomized into, and (b) that player's eventual performance in the MLB. In order to isolate the effects of norm-consistency independent of prediction accuracy, the forecasted predictions were identical for both conditions; we simply manipulated the stated source of this prediction. We note that this is a more conservative test of our hypothesis since in the Dietvorst et al. (2015) experiment, the algorithms were higher-performing than the participants. Instead, our relatively lower-performing algorithms should, if anything, be expected to engender *even less* support.

After observing the 10 forecasts from either a human or an algorithm, participants were then randomized to read a brief vignette manipulating perception of the norm followed by a set of manipulation checks. These vignettes were adapted from articles describing the prominence of either algorithm-driven management decisions (i.e., sabermetrics) or human-driven management decisions (i.e., talent scouts) for professional baseball teams. Additionally, participants were explained the incentive scheme: If selected, participants would receive 10 cents for every forecast that was within five percentile points of the player's eventual performance. Participants then made the key decision of interest: they chose to yoke their bonus to either the forecasts of the professional talent scout or the algorithm.

7.2. Results

Similar to Experiment 2, we measured the percent of participants in each condition who decided to switch decision procedures from the one they observed. As before, we treat the decision to switch from the observed procedure as a proxy for the extent to which participants penalize the human or algorithm for its observed errors. Unlike in Experiment 2, this is an actual, incentive-compatible choice rather than a hypothetical selection. As in all previous studies, higher values on this dimension corresponds with a greater penalty for a given decision procedure. Based on our hypotheses, we predicted that there will be a main effect of norm consistency on people's desire to switch, such that participants who have observed the norm-inconsistent option will be more interested in switching to the other decision procedure.



Note: Error bars represent 95% confidence intervals

Figure 2-7. *Participants' decision to tie their bonus to the procedure they observed or to switch, depending on consistency with the stated norm.*

Using logistic regression, we find a significant effect of norms when predicting the incentive-compatible binary choice of an algorithm over a human forecaster ($\beta_{\text{inconsistent}}$ = 1.89, t = 11.62, p < .001). Interestingly, for reasons that we do not observe but suspect are related to the specific domain of MLB forecasting, conditional on norm-consistency, we here observe a *preference* for algorithms ($\beta_{\text{observed\_algorithm}}$ = -0.63, t = -4.04, p < .001).

As in Experiment 1, we additionally compare the difference-in-differences between Condition A and Condition D (observing human versus algorithmic forecasts when the norm is for *humans* to forecast) versus Condition C and Condition B (observing human versus algorithmic forecasts when the norm is for *algorithms* to forecast). In Figure 2-8, we observe a similar crossover effect as in Experiment 1 (Figure 2-2): People excessively penalize algorithms compared to human forecasters when the norm is for humans to forecast, but people excessively penalize *humans* compared to algorithmic forecasters when the norm favors algorithmic forecasts ($\beta_{\text{interaction}}$ = 3.80, t = 11.50, p < .001).

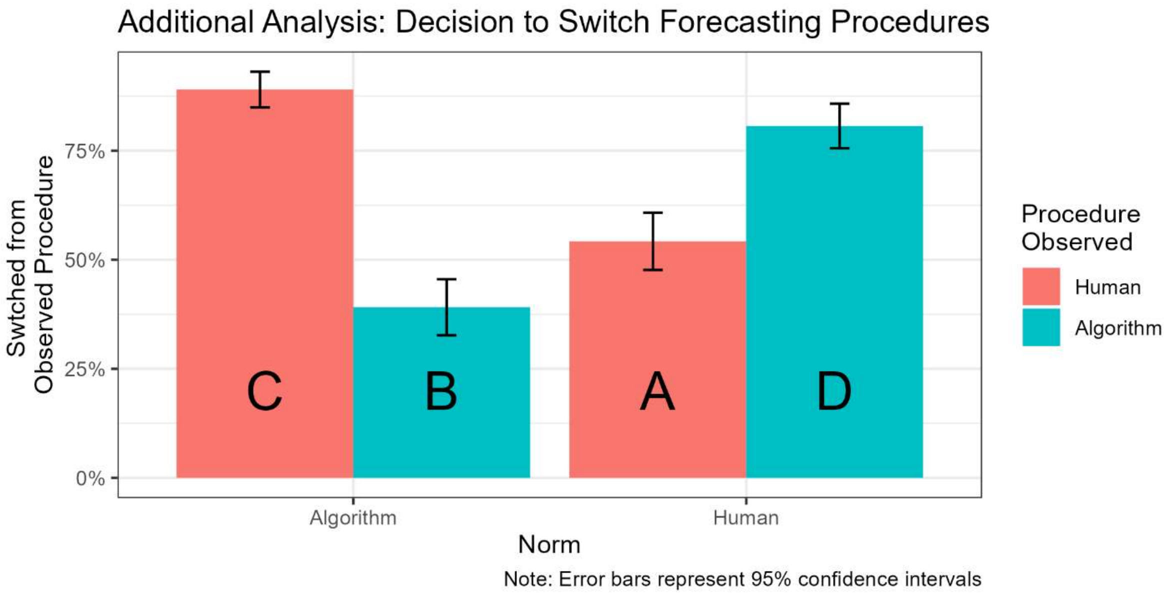


Figure 2-8. *Participants' decision to tie their bonus to the procedure they observed or to switch, depending on whether the stated norm is human or algorithmic forecasting.*

As in each study prior, we see that the influence of norms overwhelmingly shapes people's response to observing an algorithm.

7.2 Discussion

In a conceptual replication of our previous studies using an experimental design closely mirroring Dietvorst et al. (2015), we once again find strong evidence in favor of the central role of norms in AA. In particular, consistent with our Hypothesis 1, we find that observing a norm-inconsistent decision procedure—whether human or algorithm—leads to a larger desire to switch to an alternate procedure. While we expected to also observe a main effect of AA in addition to the main effect of norm-inconsistency aversion, we instead find a preference for algorithms in this study once norms are accounted for. Only when the norm favors humans to forecast does the traditional AA finding replicate.

8. General Discussion

The AA hypothesis asserts that seeing an *algorithm-generated decision* turn out badly will elicit an increased penalty (relative to humans). Norm Theory holds that seeing a *counter-normative decision* turn out badly will elicit an increased penalty (relative to norm-adhering decisions). Across five studies and several domains, we have shown that a large share of people's apparent aversion to algorithms can be explained by a more fundamental aversion to *any* decision procedure that is opposed to the norm. Thus, when the norm favors the use of an algorithmic forecaster—as is the case for myriad domains of daily living, from navigation to matchmaking—people often prefer an algorithm over a human forecaster.

People's tendency to penalize counter-normative decisions more harshly than norm-adhering decisions can easily be mistaken for veritable Algorithm Aversion any time the use of an algorithm is counter-normative. That is, what appears to be an aversion to algorithms *per se* might actually be an aversion to the algorithmic procedure being counter-normative. We suspect that these two factors have been confounded in much of the AA literature. For many of the domains in which Algorithm Aversion has commonly been studied (radiological screening, university admissions, selection of mental health providers, candidate resume screening, and more), the norm is for humans, not algorithms, to decide. If this is so, our work helps to reconcile otherwise incompatible findings of algorithm aversion in some cases (e.g., Dietvorst et al., 2015) and algorithm appreciation in others (e.g., Logg et al., 2019).

Moreover, our work extends AA scholarship by revealing a psychological mechanism that could be driving a large share of people's aversion to algorithms. This understanding can thus be used to better understand other moderators of AA. For instance, Castelo and colleagues (2019) argue that people prefer humans to algorithms for subjective tasks (e.g., predicting a joke's funniness, recommending a gift) but instead prefer algorithms to humans for objective tasks (e.g., analyzing data, predicting the weather). However, there are apparent counterexamples to their theory (e.g., driving a truck is seen as a highly objective task but people trust algorithms far less than humans; conversely, recommending a romantic partner is seen as a highly subjective task despite the popularity of algorithm-driven dating apps). If, for the tasks they chose to study, there happens to be a correspondence between the degree of objectivity and the prevalence of algorithms in those task domains, our theory can help make sense of these apparent counterexamples.

Further, our work opens up new avenues for interventions to help overcome an aversion to algorithms that might improve people's welfare: Informing people when a descriptive (Schultz et al., 2007) or trending (Mortensen et al., 2019) norm favors use of an algorithm could increase utilization. Additionally, to the extent that decision-makers interpret defaults as suggestive of a descriptive or injunctive norm in favor of the defaulted option (Everett et al., 2015; Jachimowicz et al., 2019; Mckenzie et al., 2006), setting an algorithmic forecaster as the default might itself be enough to overcome many people's aversion.

It is possible that a more general omission bias (Ritov & Baron, 1995) helps to explain some of the effect of norms in moderating AA. If deviating from a norm is seen as an action whereas adhering to a norm is seen as an omission, we would expect to see greater regret associated with any decision procedure that deviates from the norm (Kahneman & Tversky, 1982). This suggestion is similar to a finding observed by Feldman and Albarracín (2017). When algorithms are taken to be counter-normative for a particular decision, we would expect that they would be avoided since choosing them would be seen as a (more blameworthy) commission if the decision turned out poorly, whereas sticking with the norm-consistent (human) decision process would be a less-blameworthy *omission*. This would also be consistent with our findings that reversing the norm often reversed people's preferences for a human over an algorithm. Future work could more deeply investigate the question of when using a human versus algorithmic decision maker leads to higher regret and blame for poor outcomes.

One obvious limitation of the present investigation is that we have not explored how algorithms come to be accepted as the norm, simply what happens once they are. As the case of online dating offers (Rosenfeld et al., 2019), social conventions can rapidly shift to normalize the use of algorithms in a domain that was once exclusively the purview of human decision-makers. While there is likely a positive feedback cycle between the normality of an algorithm and its accuracy, this cannot be the whole explanation for acceptance since people will sometimes reject even obviously better-performing algorithms (Dietvorst & Bharti, 2020). Here, we have not explained the sociological process by which an algorithm comes to be normalized, but rather we have looked at the downstream effects of this process.

In the present investigation, in order to better understand AA, we sought to explain a critical mechanism by which AA seems to take hold. In exploring part of why AA occurs, we also sought to explain when AA can be expected and when its opposite should occur. Finally, we sought to offer a pathway for overcoming an aversion to algorithms that could improve people's welfare by highlighting norms favoring the algorithm. We suspect that the overall amount of preference or aversion to algorithms can largely be explained by people's perception of the prevailing norm for that particular domain. While algorithmic forecasting offers great promise for improving people's lives, they are useless if people refuse to heed their recommendations. Understanding why people are averse to algorithms can reveal new insights from their study and new pathways for their adoption.

Appendix

Experiment 2 Additional Analyses

1. Continuous Versus Binarized Dependent Measure of Choice

In the manuscript, for the sake of rhetorical fluency, we binarize the dependent measure of interest. In fact, participants rated their decision to use a human or algorithm using a seven-point scale (-3 = definitely human, +3 = definitely algorithm). In the manuscript, we treat anyone who selected a negative number as having chosen a human and anyone who selected a positive number as having chosen an algorithm. Instead, we could replicate the analyses from the manuscript taking advantage of the full variation of the continuously measured Likert question as it was asked.

The following table documents the results of modeling the choice of an algorithm from an interaction between observing an imperfect human with the norm-consistency of this procedure using either a binary (Model 1, as reported in the manuscript) or continuous (Model 2, as asked) choice measure.

Table A2-1

	Dependent variable:	
	Binarized	Likert
	(1)	(2)
Observed human	0.718^{***}	3.199^{***}
	(0.053)	(0.220)
Norm consistent	0.403^{***}	1.907^{***}
	(0.050)	(0.212)
Obs human X Norm consistent	-0.879^{***}	-3.738^{***}
	(0.075)	(0.310)
Intercept	0.037	-2.222^{***}
	(0.035)	(0.151)
Observations	473	548
R^2	0.291	0.286
Adjusted R^2	0.286	0.282
Residual Std. Error	0.405 (df = 469)	1.809 (df = 544)
F Statistic	64.097^{***} (df = 3; 469)	72.729^{***} (df = 3; 544)
Note:		*p<.05, **p<.01, ***p<0.001

As can be seen, while the scaling of the variables changes between models—as well as the interpretation of the coefficients—the significance testing remains qualitatively identical.

2. Equivalencies of Dependent Measures and Predictors

There are two logically equivalent ways that we can think of the dependent measure of interest: (1) The share of participants who chose a decision procedure different from the one that they observed (i.e., “penalizing the observed procedure”), or (2) The share of participants who chose to utilize the algorithmic decision procedure, conditional on whether they observed a human or algorithmic decision procedure. For rhetorical fluency, in the manuscript we presented our analyses in the former way. As is clear upon reflection, these two measures are the same when the observed procedure is the human and complementary when the observed procedure is the algorithm.

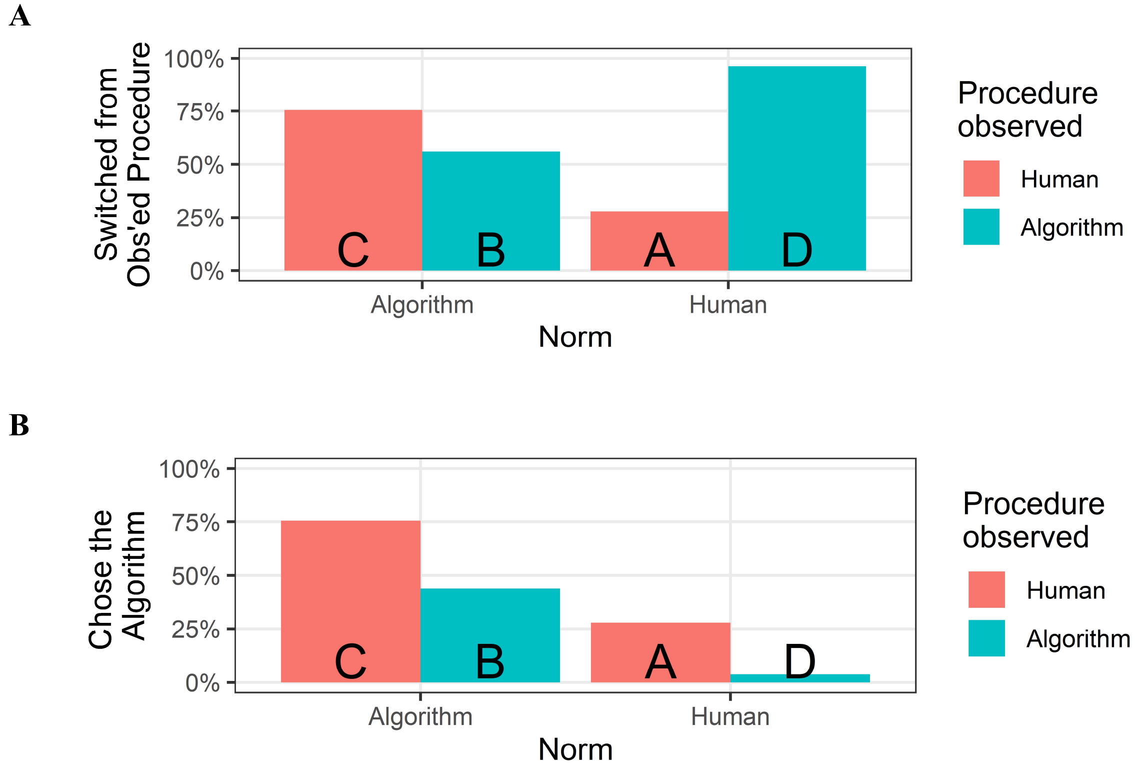


Figure A2-1: *(A) Penalty of the observed procedure, depending on norm (human vs algorithm), and (B) An extensionally equivalent representation of the same data, where Bar B and Bar D from Figure (B) are complements of the corresponding bars in Figure (A).*

Comparing the graphs above makes this point clear. Whenever the observed procedure was a human (Bar C and Bar A), the dependent measure is the same using either description (choice of algorithm or decision to switch procedures). Whenever the observed procedure was an algorithm (Bar B and Bar D), the dependent measures are complementary (i.e., Percent choosing the algorithm = 100% - Percent switching from the observed procedure). In this way, the two ways of describing the results are fundamentally the same. This is similar to the relationship described in the paper between describing the results in terms of *a norm of using an algorithm* (whether observing a human or algorithm) versus *the consistency of the observed procedure with the stated norm*. While interpretation changes, the results remain logically synonymous.

3. Additional Analyses

In the manuscript, we note that the results of Study 2 extend Study 1 by considering a choice-based DV. As mentioned, we also asked participants about expected regret as a mediator. As noted, in addition to finding mediation by expected regret, we also replicate the effects on expected regret as in Study 1. The following table summarizes these findings.

Table A2-2

	Dependent variable:	
	Expected regret	
	(1)	(2)
Observed algorithm	0.797^{***}	0.078
	(0.157)	(0.223)
Norm inconsistent	0.731^{***}	
	(0.157)	
Norm human		-0.546^{*}
		(0.229)
Observed Alg X Norm human		1.441^{***}
		(0.314)
Intercept	3.237^{***}	3.874^{***}
	(0.138)	(0.163)
Observations	548	548
R^2	0.080	0.082
Adjusted R^2	0.077	0.077
Residual Std. Error	1.837 (df = 545)	1.836 (df = 544)
F Statistic	23.782^{***} (df = 2; 545)	16.274^{***} (df = 3; 544)
Note:		*p<.05, **p<.01, ***p<0.001

As can be seen, the results from Study 1 are replicated using the regret-based dependent measure in Study 2.

CHAPTER 3

The Choice Architect Doth Protest Too Much:

Ironic Effects of Nudging on Perceptions of Descriptive Social Norms

Noah J. Goldstein
University of California, Los Angeles

Jonathan E. Bogard
University of California, Los Angeles

ABSTRACT— Behavioral science has achieved notoriety for offering scientifically informed low-cost interventions that “nudge” people toward good behavior. When implemented, however, many nudges have failed to produce results found in the lab. These failures are often attributed to problems related to unobserved moderators, the replicability of the original finding, or errors implementing or scaling the intervention. In the present paper, however, we point to a different culprit: The mere presence of a nudge can “leak information” about negative antecedent circumstances that brought about the decision to implement the nudge. Specifically, across three studies and a variety of domains, nudges, and judgments, we show that various behavioral interventions can lower people’s perceptions of the descriptive social norm favoring the desired behavior. This implicit negative social proof, we conjecture, may contribute to failures of nudge implementation. By documenting this ironic effect of nudging, we hope to deepen our theoretical understanding of behavioral science in the wild.

Keywords: sensemaking, choice architecture, nudging, ironic effects, descriptive social norms

In the early 1990s, the government of Switzerland sought to create new repositories for nuclear waste containment. The government's first step was to call for a national referendum to determine which cantons would house the repositories. Nuclear waste containment sites pose health risks to nearby residents and also lower local property values, so the decision of where to locate the facilities was contentious. During this time, Frey and Oberholzer-Gee (1997) surveyed Swiss citizens living in cantons being considered for siting the repositories. Perhaps amazingly, more than half of all respondents supported allowing the nuclear waste repositories to be built in their local community. However, when respondents were additionally offered a considerable annual monetary compensation for allowing the sites to be built nearby—approximately equivalent to six weeks of the mean Swiss salary—support *fell* by about 50%.

Adding inducements for a desired behavior can often, ironically, lead to less of that behavior (Benabou & Tirole, 2003). One critical reason for this may have to do with the negative inferences that people make from the mere presence of the inducement. For example, Benabou and Tirole show that principal–agent incentive schemes can be interpreted as signals of both the principal's distrust of the agent as well as private information about the unpleasantness of the task. On either interpretation, agents make negative inferences in response to the simple presence of an incentive. In another demonstration of this phenomenon, Cryder and colleagues show that people often assume that research studies offering larger compensation come with commensurately greater personal risk from participation (Cryder et al., 2010). Relatedly, awareness of the incentives that others face has been shown to lower take-up of recommendations (Verlegh et al., 2013) and to decrease reciprocation of prosocial behavior (Orhun, 2018) due to negative inferences about others' motives.

Sometimes these negative inferences can lead to unexpected outcomes. For instance, when Ugandan health-promoters posted job solicitations offering an unusually high salary, the eventual performance and retention of people hired under this regime was worse than when the job paid a lower salary (Deserranno, 2019). Explaining this, a survey of applicants revealed that when the salary was higher people assumed that the job was less socially beneficial. This inference, in turn, dissuaded pro-socially motivated people from applying. In another study of the surprising effects of economic inducements, Fehr and List (2004) modified a Trust Game to include a condition in which participants who shirk can be punished. They find that the possibility for one player to punish another is enough to provoke overall less trusting behavior in participant interactions. Apparently, the threat of punishment communicated a negative expectation of how trustworthy the other player would act.[10] In each of these cases discussed above, awareness of an economic inducement caused people to make unfavorable inferences about the background circumstances that gave rise to the inducement. In other words, people often presume that incentives typically signal the *need* for an incentive, which itself is a signal of negative baseline circumstances.

In such situations, the incentives can be said to "leak information" about the background context, including the beliefs and intentions of the policymakers (henceforth, "choice architects," people who influence the design of choice environments). The process of interpreting signals from the context (above, for example, the economic incentive schemes) is known as "social sensemaking" (Weick, 1995). When sensemaking, people look to the features of a situation (the incentives, structures, options, hierarchies, characterizations, and so on) as meaningful

[10] Of note, however, if players signaled mutual trust by voluntarily eschewing the shirking penalty, this led to an *increase* in trust behavior from their partner.

information about the underlying intentions and beliefs of the choice architect as well as the prevailing circumstances that gave rise to specific decisions of the choice architect. Similarly, people use cues from their social interactions (Gilbert et al., 1988; Grice, 1975; Schwarz, 1994) and environment (Kamenica, 2008; Weick et al., 2005; Wernerfelt, 1995) to make meaning of the situation. Whenever people use the presence of an economic incentive to infer something negative about the background circumstances, they are using the incentive scheme as a meaningful input to social sensemaking.

Economic inducements can be thought of as just one example of a more general class of interventions aimed at encouraging a particular behavior. The field of behavioral science has studied myriad other, non-economic inducements of behavior, and documented their impact on people's decisions. This invites the intriguing question of whether these other forms of choice architectural interventions ("nudges") also leak information about the background circumstances similar to the way that economic inducements do. That question is the subject of the present investigation.

A growing body of literature has shown that social sensemaking is sometimes engaged by the presence and design of nudges. For example, people often infer that a particular option is being recommended by the choice architect if it is set as the default (Dinner et al., 2011; Jachimowicz et al., 2019; Mckenzie et al., 2006). Similarly, listing minimum repayment amounts on credit card bills may be interpreted by debt holders as *suggested* repayment amounts (Navarro-Martinez et al., 2011). Even more germane to the present investigation, a working paper by Tannenbaum and colleagues argues that another feature of choice architecture, menu partitioning, can tacitly suggest to people that the more finely partitioned categories contain the most popular options (Tannenbaum et al., 2017).

Related work in marketing research has documented similar effects. For instance, Brown and Krishna (2004) find that when “marketplace metacognition” (i.e., social sensemaking) is triggered, depending on its interpretation, a default policy can actually backfire. Relatedly, despite their popularity among marketing consultancies, inspirational and values-based persuasion appeals often backfire because targets infer from this tactic that the marketer has ulterior motives (Alavi et al., 2018). More generally, when consumers find persuasive appeals or sales tactics inappropriate (Campbell & Kirmani, 2000; Friestad & Wright, 1994), thereby triggering social sensemaking (Wright, 2002), they may react against the persuasion attempt. In this way, if social sensemaking is engaged, people may spontaneously make negative inferences about the background circumstances that gave rise to the intervention.

This work is drawn together in a recent framework proposed by Krijnen et al. (2017), known as “Choice Architecture 2.0.” This model construes choice architecture as an implicit conversation between the choice architect and the decision maker. The authors suggest that features of the choice environment can signal information about the choice architect’s beliefs and intentions, and that these signals are then taken as useful information for the decision makers. Becoming aware of a choice architectural intervention may cause decisions makers to wonder about *why* the intervention was put in place. They may treat the existence of the intervention as a meaningful signal about both (a) the background context that begat the intervention, and (b) the beliefs and intentions of the choice architect that led them to implement the intervention.

Here, we hypothesize that people may reason that one plausible justification for creating an intervention is the choice architect’s concern that not enough people are choosing the desired behavior. Concretely, we expect that awareness of a nudge can cause people to think, “Someone designed it this way to get me to [X]. They were probably worried that not enough people were previously doing [X]" (where "[X]" is any behavior intended by a nudge, such as increased retirement savings, vaccine take-up, exercise, and so on). In other words, people may infer from the mere presence of a nudge that the promoted behavior is currently unpopular. Put otherwise still, a solution may imply a problem.

The Present Paper

Putting this together, we expect that when people become aware of a nudge[11] and social sensemaking is triggered, they will often infer that the nudge was intentionally installed in response to a problem. Further, we expect that this problem will often be assumed to be a descriptive social norm (Cialdini & Trost, 1998) opposed to the behavior intended by the nudge. To illustrate this process concretely, consider a person who encounters a novel message on their tax form that reads, "Did you know that the majority of people in your zip code file their taxes by April 15?" Because this message is unexpected, this citizen may wonder why the message was introduced (Maitlis & Christianson, 2014). One plausible inference is that this was the government's attempt at encouraging even more people to file their taxes promptly. Further consideration might suggest that the government is seeking greater compliance because they are unhappy with the currently low rates of on-time filing. If this is right, the mere presence of the nudge would be enough to imply a lower-than-desired descriptive social norm of timely tax filing. While we expect this kind of process to unfold for a variety of nudges, we pause to note the ironic effect of this particular kind of nudge as it is a "norms nudge," designed to encourage good behavior via social proof (Bicchieri & Dimant, 2019).

[11] Defined informally here as any kind of intervention on the choice context, designed to promote a certain desired behavior, without changing the underlying option set or financial incentives. For a more rigorous treatment, see Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: Improving decisions about health, wealth, and happiness*.

We test this proposed effect across a range of domains, decisions, and nudges in three pre-registered studies. We begin with Study 1 by testing the hypothesis that a nudge can lead to negative inferences about the perceived descriptive social norm ("PDSN") of the desired behavior. Study 2 replicates this finding in a new domain and demonstrates a central mechanism: perceptions that the choice architect is worried about the situation. In Study 3, we show further evidence of this mechanism through moderation: By providing an alternative justification for the nudge that does not bear on the level of choice architect concern, we eliminate the effect on PDSN. All open-science materials (pre-registrations, study materials, data, and code) can be found online at https://researchbox.org/685.

Study 1: Ironic Effect of Norms Nudges

We hypothesized that there can be a negative effect of nudges on people's perceptions of the antecedent PDSN favoring the desired behavior. If that nudge is a *norms nudge* (i.e., a social proof intervention), this would be especially ironic since this negative effect on the PDSN would directly countervail the intended direct effect of the nudge itself. To test this prediction, we compared the effects of a norms nudge against both a passive control (i.e., compared to baseline beliefs) and against the effect of learning the same information contained by the norms nudge through a different channel.

Method

Participants. In Study 1, we recruited a convenience sample of 753 American participants[12] from Cloud Research's "approved participant" pool of Amazon's Mechanical Turk (MTurk) workers with a 95% MTurk rating or higher. Subjects were included only if they had participated in fewer than 500,000 total surveys and if they had not participated in a related research study from our research team. Per our pre-registration plan, we excluded anyone who failed an initial attention-based screener, a test/retest consistency check on their birth year, or an instructional manipulation check resembling the key measure (Oppenheimer et al., 2009). This screening procedure was consistent for all studies reported in the present paper. After exclusions, we ended with a final sample of 725 participants (44% female, M_{age}=40.2, SD_{age}=12.4).

Materials and Procedure.

Following the initial screener and consent form, participants were asked to imagine that they wanted to sign up for a dating site and had begun creating their online profile. Participants were then asked to imagine that, after answering several basic background questions, they came to a field where they were to input their weight. Participants were then randomized into one of three experimental conditions:

[12] We targeted a sample of 750 participants but there is some imprecision in Cloud Research's recruitment procedure. Hence, as in all studies for this paper, we ended up with a few participants more than our targeted number.

- **Treatment**: Participants saw a screenshot of a popup informing them that the majority of users honestly report their true weight.
- **Informational Control**: Participants were told that they were curious about honest reporting on dating websites so they Googled for information. They were asked to imagine finding an interview of an employee of the app they were using. The interview included the employee's claim that the majority of users honestly report their true weight.
- **Passive Control**: Participants were given no additional information.

Of note, those in the Treatment and Informational Control conditions saw an identically worded claim—"According to our data, the majority of our users are completely honest in reporting their actual weight"—only the context of encountering the claim varied (popup versus article).

Following this, participants were asked on a seven-point scale how common they think it is for users of the website to give false information about their weight (1=Not at all common, 7=Extremely common), then were asked an estimate of the percent of users who honestly report their true weight when filling out their profiles (0-100). The former question can be thought of as participants' general sense of how big of a dishonesty problem the community faces; the latter is a direct measure of the PDSN of honesty on the app. We expected that, compared to encountering the same information in a different channel, encountering the claim via a norms nudge would lead to a significantly lower PDSN of honesty (and thus a greater problem of dishonesty). As an exploratory analysis, we were curious to see if this expected backfire from nudging the information would lower the PDSN entirely back to baseline (i.e., the Passive Control) or if people would still update their beliefs to some extent in the direction of the normative information.

Results

The purpose of Study 1 was to test whether a norms nudge can have the ironic effect of decreasing the PDSN compared to learning that same information through a different channel. Indeed, that is what we find (b=-6.16, t(719)=-2.94, p=.004).[13]

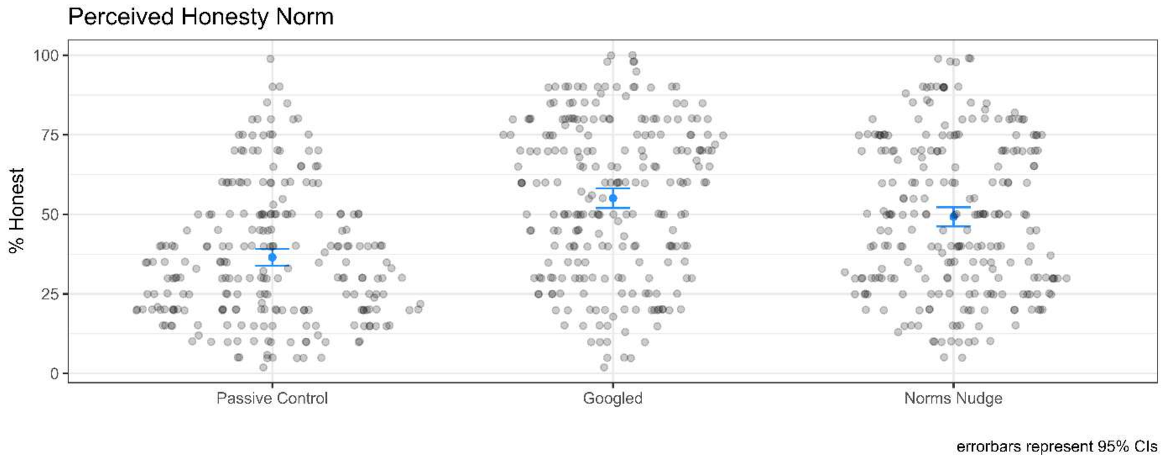


Figure 3-1. *Participants' judgments of the percentage of dating app users who honestly report their weight, depending on whether they received no information, learned from reading a news article that the majority of users* do *honestly report their weight, or learned this same information from an onscreen popup (i.e., nudge) while filling in the dating profile.*

Relatedly, the norms nudge also led to considerably greater perceptions of the problem magnitude using the Likert measure of dishonesty (b=0.52, t(719)=3.51, p<.001). In addition to the relative effects, the absolute levels of each condition's means are interesting to consider (Figure 3-1). At baseline, people assume that the majority of all users are dishonest (i.e., a negative PDSN). Independently learning that "the majority" of users are honest boosts the PDSN of honesty to approximately 55.1% (SD = 24.4). However, learning that the majority are honest instead through a norms nudge yields a mean PDSN of honesty hovering right around 50% (M=049.2, SD=24.0). This suggests that many who encounter this norms nudge may believe that

[13] See the Appendix for full regression tables of all studies in the present paper.

the message could be approximately true, *but just barely*. In sum, we find evidence consistent with the hypothesis that, compared to Googling, learning normative information via a norms nudge leads to a considerably lower PDSN and considerably higher judgment of the underlying problem.

We were further interested in the question of how large the backfiring effect of the norms nudge was relative to baseline beliefs. To test this, we compared the PDSN for those who received the norms nudge to those in the passive control condition, whose baseline beliefs were estimated absent any normative information. We find that participants estimated a significantly greater social norm if they encountered the nudge compared to those who received no normative information (b=12.63, t(719)=6.04, p<.001). Results were similar for perceptions of the magnitude of the problem (b=-0.79, t(719)=-5.41, p<.001). Put otherwise, while the norms nudge backfired with respect to independently learning the normative information, it still had a positive effect relative to baseline beliefs. This is consistent with the broader literature on norms interventions, which has often demonstrated success in shifting people's beliefs and actions in the desired direction (e.g., Bogard et al., 2020; Goldstein et al., 2008; Schultz et al., 2007). Taken together, these results suggest an interesting picture of normative interventions. There seems to be two components of norms nudges: the *norms* part (i.e., social proof), which can positively affect beliefs and actions, and the *nudge* part (i.e., an intervention intentionally installed by some individual) which can arouse suspicions about an antecedent normative problem. The relative magnitude of these two vectors, we suspect, might explain a critical difference between prior studies that have successfully and unsuccessfully attempted to change behavior using normative information nudges.

Study 2: Mechanism: Inference About Choice Architect Concern

In Study 1 we found that perceptions of the norm are lowered when normative information is conveyed via nudge rather than learned through an independent channel. Relatedly, people perceive a greater underlying problem of negative behavior. These are the conditions that might plausibly give rise to a choice architect implementing a nudge in order to improve the situation. The purpose of Study 2 is to test this claim. Here, we evaluate the hypothesis that implementation of a nudge lowers the PDSN in part due to its suggestion that the choice architect is worried about the current descriptive social norm. Additionally, another purpose of Study 2 was to test the generalizability of our findings from Study 1 in a new domain (self-reported academic credentials on a job-search website). The social meaning, norms, verifiability of information, baseline rates of honesty, and other critical features of this domain differ importantly from filling out a dating profile, so we thought this would be another informative, familiar domain in which to test our hypotheses. Finally, given a convenience sample of online workers, we thought that measuring the effects of interventions to promote honesty in an online job-search website would be especially interesting.

Method

Participants. In Study 2, we recruited 761 participants from MTurk using our standard inclusion criteria (see Study 1). We were left with a final sample of 733 participants (51% female, M_{age}=42.9, SD_{age}=12.9).

Materials and Procedure.

The structure of Study 2 was quite similar to that of Study 1. Following the initial screener and consent form, participants were asked to imagine they were interested in finding a job by joining a (fictional) job-search service online (TalentMatchers). We asked participants to imagine that, after joining the website and answering several basic background questions, they were then asked to input the GPA from their highest-awarded degree (GED, High School, Associates, Bachelors, and so on). Participants were then randomly assigned into Treatment, Informational Control, or Passive Control conditions as in Study 1. This time, the information received—either as a popup (Treatment) or through independent research (Informational Control)—claimed that the majority of TalentMatchers users honestly report their true GPA. After that, unlike in Study 1, participants then answered a question about the extent to which they thought TalentMatchers was troubled by the level of dishonesty on their website. After this, similar to Study 1, participants then indicated their estimate of the level of honest GPA reporting on the TalentMatchers site (i.e., the PDSN).

Results

The central purpose of Study 2 was to replicate the findings from Study 1 and to test whether this phenomenon is partially driven by perceptions of choice architect concern. We find evidence consistent with both predictions.

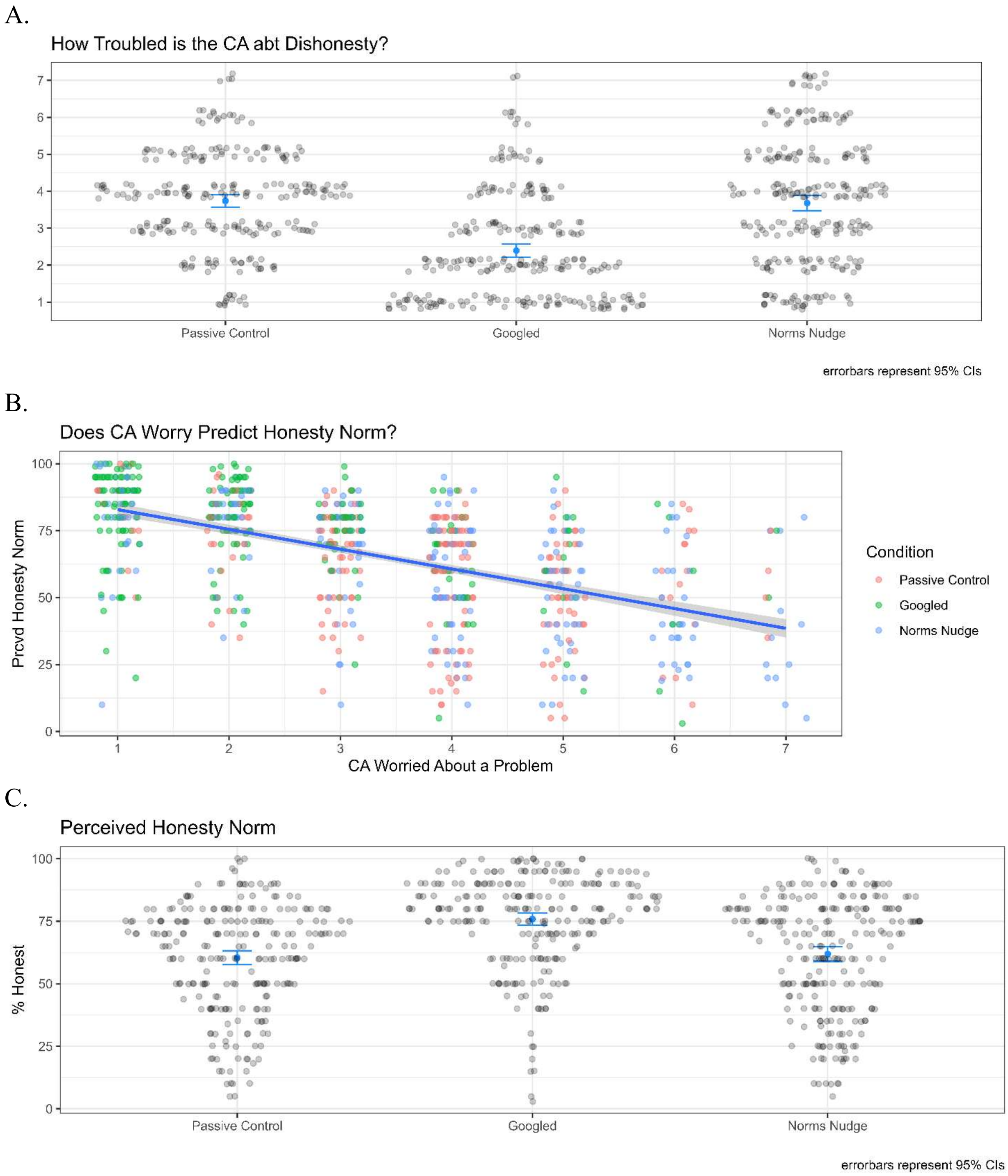


Figure 3-2*. (Fig. 3-2A) A-Path Effects: Ratings of how troubled the choice architect (TalentMatchers) likely is about the level of dishonest reporting on their site, by experimental condition. (Fig. 3-2B) B-Path Effects: Correlation between perceptions of choice architect worry and PDSN of honest reporting. (Fig. 3-2C) Total Effect: PDSN of honest reporting, by experimental condition.*

As before, we find that learning normative information from a nudge, compared to learning the same information from an independent information channel, lowers the PDSN (b=-14.02, t(729)=-7.30, p<.001; Figure 3-2C). Thus, we replicate the findings from Study 1 documenting an ironic backfiring effect of norms nudges on perceptions of the norm.

We were also interested in testing the proposed mechanism: perception of choice architect concern. Here we find that, compared to the Informational Control (i.e., Googled condition), the norms nudge led to significantly greater perceptions that the choice architect was troubled by the amount of dishonesty on their site (b=1.28, t(729)=9.52, p<.001; Figure 3-2A). Further, we find a strong negative correlation between perceptions of choice architect concern about dishonesty and the PDSN of honest reporting on the website (b=-7.39, t(731)=17.08, p<.001; Figure 3-2B). We used 10,000 simulations[14] to bootstrap 95% confidence intervals for a statistical mediation model.

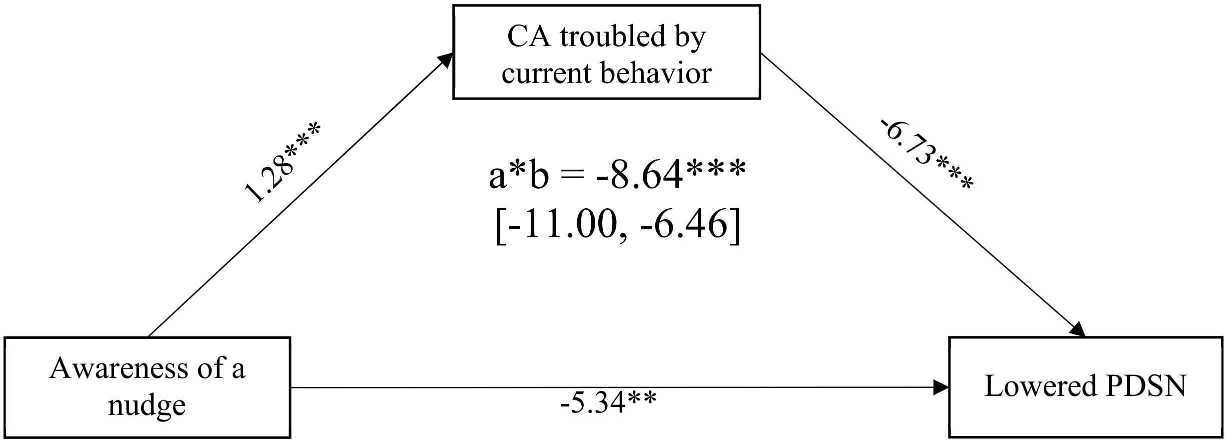


Figure 3-3. *Mediation model estimating the indirect effect of the presence of a nudge on the perceived descriptive social norm (PDSN) via the belief that the choice architect is troubled by the current behavior. Note: "CA" stands for "Choice Architect."*

[14] Seed: 112620

We find evidence of a significant indirect effect of the norms nudge on PDSN *through* perceptions of choice architect worry (b_{med}=-8.64, 95% CI [-11.00, -6.46]). We thus take this as support for the idea that part of what drives the backfiring effect of norms nudges on PDSN is the information leaked from the mere presence of the intervention regarding the thoughts and intentions of the choice architect.

As before, it is worth commenting not just on relative differences between conditions but also on absolute levels of the PDSN of honest reporting in this domain. Unlike in Study 1, the baseline (i.e., Passive Control) PDSN of honesty here was greater than 50% (M=60.5, SD=21.6). This creates an interesting context for testing our effect since the normative information provided to the other two conditions claimed only that a "majority" of users honestly report their GPA. While learning the normative information from an independent channel led to a 25% increase in the PDSN compared to baseline (M=75.9, SD=18.8), *nudging* this same information left people with a PDSN statistically indistinguishable from those who received no normative information at all (M=61.9, SD=23.2; t(729)=0.74, p=.46 *NS*). To test this relationship, we used Bayesian estimation of the probability of the null hypothesis—that there is no difference between a norms nudge and the baseline beliefs from the passive control—conditional on the data that we observe. Taking pains to not construe this as evidence that the two conditions are in fact equal, we find the null hypothesis to be very likely ($P(H_0|Data)$=.954, Bayes Factor=20.64) given the data observed (Wagenmakers, 2007). In summary, we once again see that the mere presence of a norms nudge, compared to learning the same information independently, lowers the PDSN of the desired behavior. In fact, the backfiring effects of the nudge seem approximately as potent as the positive effects of the normative information it contained. Further, we find evidence that this may result from perceptions that the choice architect is troubled by the baseline levels of negative behavior. This worry, presumably, is what causes the choice architect to implement a behavioral intervention (i.e., nudge). In this way, the nudge itself may be communicating a certain "negative social proof" that is undermining to its very purpose (Cialdini, 2021; Cialdini et al., 2006).

Study 3: Mechanism by Moderation

There were three distinct goals of Study 3. First, chiefly, we wanted to test for further evidence of the proposed mechanism (perception that the choice architect is worried about the current behavior). Given the limitations of statistical mediation as a test of causal mechanism, we sought to bolster the evidence from Study 2 by testing our proposed model through moderation. Thus, in Study 3 we seek to block the pathway from *presence of the nudge* to the *inference of choice architect concern* (i.e., the a-path; Figure 3-3). Blocking this inference ought to attenuate the total effect of nudging on lowered PDSN if our model is correct. In service of this, we added a condition in which participants were given an alternative justification for the nudge's presence, one that implies nothing about the level of choice architect concern. We conjectured that this would diminish the effect of the nudge on the PDSN.

The second goal of Study 3 was to further test the generalizability of the phenomenon under investigation. While Study 2 varied the domain from Study 1, the particular behavioral intervention (a norms nudge) and the norm in question (honest self-reporting of information) were nearly identical. In Study 3, we expand into yet a different domain, looking at yet a different behavior (speeding). Further, in Study 3 we also implement a different kind of nudge (a speed radar sign) and solicit judgments of a different norm (breaking the law).

The third goal of Study 3 was to rule out an alternative explanation of our results. One interpretation of our findings so far might go as follows. People in our studies are simply acting as good Bayesian updaters—when they receive some information about the norm, they update their beliefs in the direction of this new information, to some extent, depending on the credibility of the information source. When that source is reliable (as with an independent news outlet), we observe considerable belief updating. When there are reasonable concerns about the motives of the information source (as when the owners of the app created the popup), there is good reason to update less (Study 1) or not at all (Study 2) in light of the new information. Maybe, this argument goes, people are doing nothing more than updating their beliefs precisely as we would expect Bayesian reasoners to do.

On one hand, we are sympathetic to this account of what we have demonstrated. The noteworthy finding, we think, is the fact that the simple act of providing this normative information as a nudge may be enough to engender disbelief about its veracity. In other words, the novel finding is the demonstration of unfavorable social sensemaking from the mere presence of a nudge. Beyond the theoretical insight, this also has considerable practical implications for would-be choice architects trying to nudge behavior in the field.

On the other hand, we do not think that this is the entire story. Our hypothesis holds for a wide range of nudges, including those that do not rely on belief updating for their efficacy. In such cases where the nudge does not operate by offering information (to be rejected or accepted), our findings could not be explained as apt Bayesian reasoning. Thus, in Study 3 we use a different kind of nudge—importantly, not a norms nudge—to test our general hypothesis. We predicted that, despite the very different kind of behavioral intervention, the mere presence of the nudge would be enough to lower perceptions of the descriptive social norm. In so doing, we sought to demonstrate that nudges can have the proposed deflating effect on the PDSN even when they do not operate through belief updating.

Method

Participants. In Study 3, we recruited 758 participants from MTurk using our standard inclusion criteria. We were left with a final sample of 730 participants (54% female, M_{age}=41.8, SD_{age}=13.4).

Materials and Procedure.

Following the initial screener and consent form, participants were asked to imagine that they drive along the same local streets on their commute to work every day. Participants were then randomized into one of three experimental arms:

- **Control**: No further information was given.
- **Nudged**: Participants were asked to imagine that one day, while driving to work along the usual route, they saw a speed radar sign for the first time along the route.
- **Random**: Participants were told the same thing as those in the Nudged condition. They were then told that this did not surprise them since they had heard about this program on the local news a few weeks earlier. In the news program, they learned that: (a) the neighboring county had extra radar signs and donated the surplus to their county, (b) the sheriff's office used a computer program to randomize the streets where the signs would be placed, and (c) the signs would remain in one set of streets for a few days, then the program would randomly assign them to another set of streets.

For clarity, participants in both the Nudged and the Random condition were shown a picture of a speed radar sign below the text. As in Study 2, participants were then asked to rate the extent to which they thought local law officials were troubled by the amount of speeding, and then they were asked the percent of all people who typically speed along the local streets each day. For participants in the Nudged and Random conditions, we clarified that we were asking for participants' judgments about the time *before* the signs went up to ensure that we were not eliciting their judgment of how effective the signs are at reducing speeding but instead asking about the background conditions before the signs went up.

Results

Beyond generalizing our findings, the purpose of Study 3 was to (a) demonstrate that the backfiring effects of nudging hold for behavioral interventions besides norms nudges, and (b) offer further evidence of our proposed mechanism. We offered an explanation for the presence of the radar signs that could not plausibly be suggestive of local officials' concerns about speeding. We expected that doing so would diminish the effect of the nudge on lowering perceptions of the norm. In other words, by inhibiting the a-path of our proposed mediation model (Figure 3-3), we expected to diminish the central phenomenon we have observed. This is what we find.

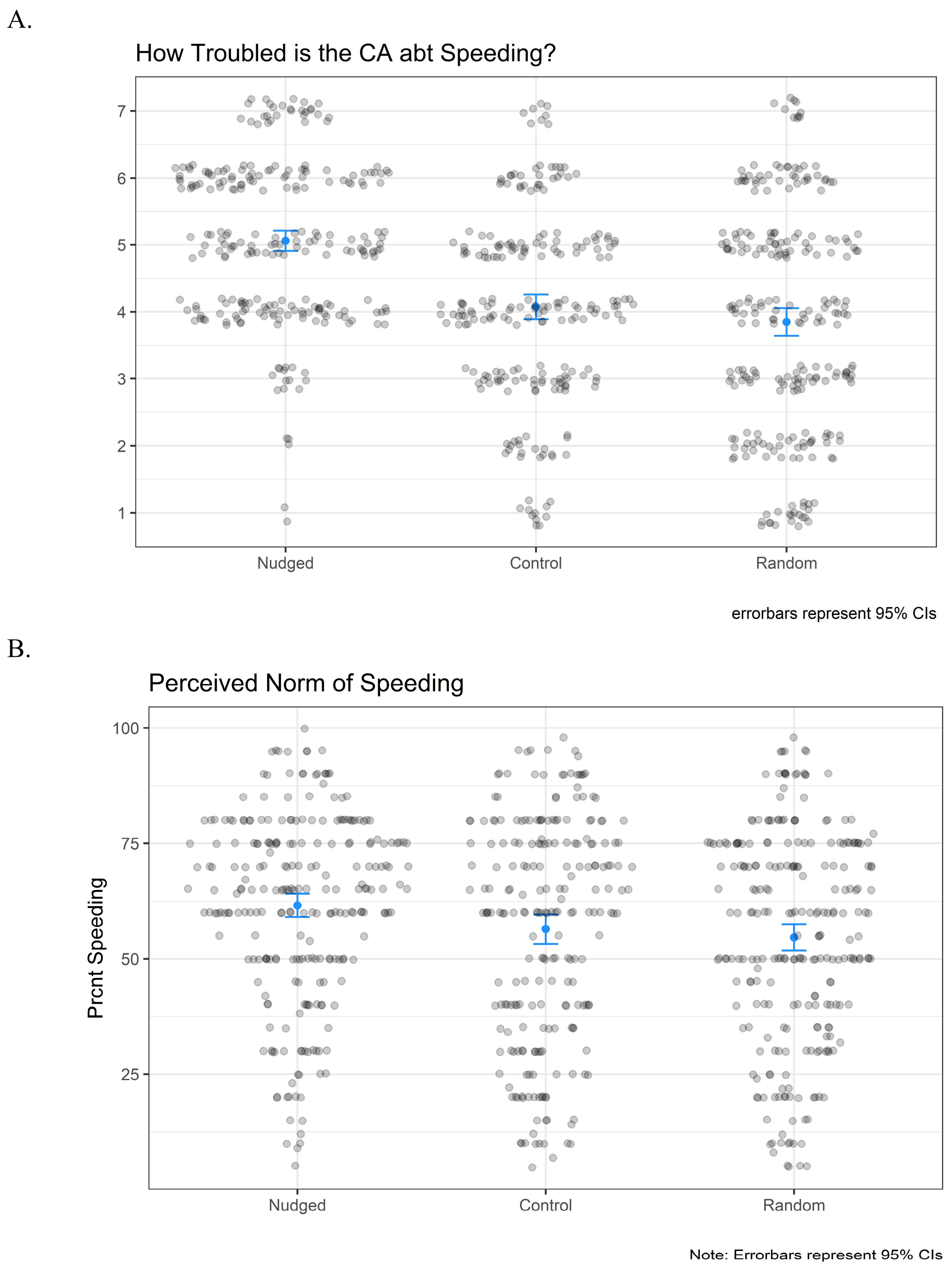


***Figure 3-4**. (Fig. 3-4A) Effect of experimental condition on perceptions that local law officials are concerned about the level of speeding along local roads. (Fig. 3-4B) Perceived descriptive social norm of speeding along local roads, by condition.*

First, we find that those in the Random condition inferred much less concern from local officials compared to those in the Nudged condition (b=-1.21, t(725)=-9.42, p<.001; Figure 3-4A). In other words, we successfully blocked the a-path inference of our model. Next, critically, we find that the PDSN of speeding is lower for the Random condition than the Nudged condition (b=-6.76, t(725)=-3.68, p<.001; Figure 3-4B). In fact, the mean PDSN of those in the Random condition was statistically indistinguishable from but directionally *lower* than the mean PDSN of people's baseline (Control) beliefs (b=-1.21, t(725)=-.64, p=.52 *NS*; Figure 3-4B). A linear hypothesis test confirmed that the difference between the Control and the Random conditions' PDSN was significantly smaller than the difference between the Control and Nudged conditions (F=13.52, p<.001). Thus, by providing a justification for the signs that did not bear on the degree of officials' worry about speeding, we successfully inhibited the effect on PDSN.

We hasten to point out further that when no such alternative justification was provided, as in the Nudged condition, we once again observe a more negative PDSN compared to Control (b=5.55, t(725)=2.95, p=.004; Figure 3-4B). This is noteworthy because it means that we were yet again able to replicate the central hypothesis regarding backfiring inferences about the PDSN. Despite using a very different sort of behavioral intervention than the nudges examined in Study 1 and Study 2, the mere presence of the nudge was enough to lower perceptions of the descriptive social norm. Further, Study 3 gives credence to the notion that providing an alternative justification for the presence of a nudge—one unrelated to choice architect concern—eliminates the backfiring effect. Piecing this all together, we take this to be corroborating evidence in support of our proposed mechanism. We also take this to be a repudiation of the claim that "all that's going on" with our findings is proper Bayesian updating in response to relevant information of varying credibility. We see that the negative effects of nudging on the PDSN extend beyond norms nudges.

Discussion

Imagine touring the basement of a house you were considering purchasing and noticing patchwork on an exterior wall and discovering a dehumidifier in the corner. From these clues it seems natural to assume that there had previously been water problems in the basement. Analogously, we have proposed that awareness of an intervention in a decision environment can serve as a meaningful clue about the *need* for the intervention in the first place. In this way, introducing a new policy (e.g., increasing fines for non-payment of taxes; only permitting one student to use the restroom at a time during an exam; adding security cameras to the employee supply closet) may cause people who encounter the policy to suspect that (a) there had previously been some problem (e.g., it was common for people to evade taxes; cheat on exams; steal office supplies), and (b) a policymaker reflectively decided to implement the given policy in order to curb that problem. We show that this holds for various behavioral interventions. While our hypotheses are not exclusive to situations in which a new policy is introduced, we only expect to observe these effects when there is (a) (conscious) awareness of the nudge, and (b) social sensemaking about the situation (conscious or not). The introduction of a new policy seems to be one natural trigger of social sensemaking (see Centraal Bureau voor de Statistiek, 2017, as discussed in Krijnen et al., 2017). Ascertaining other features besides novelty that may lead to this response (e.g., surprise, appropriateness, uncertain motives, baseline distrust, and so on) is reserved for future work (for related ideas, see Wright, 2002).

Across various domains, different kinds of behavioral interventions, and regarding judgments of different normative behavior, we have demonstrated that the mere presence of a nudge can cause people to infer that the descriptive social norm opposes the behavior intended by the nudge. Moreover, we showed that this is because people interpret the presence of a nudge as an indication that the choice architect had been troubled by the prevailing norm. This, presumably, is taken as the reason for the choice architect's deliberate decision to implement the nudge (in order to increase the positive norm). Figure 3-5 offers a conceptual model of the proposed inferential chain, from awareness of the presence of a nudge down to the belief about a lowered descriptive social norm.

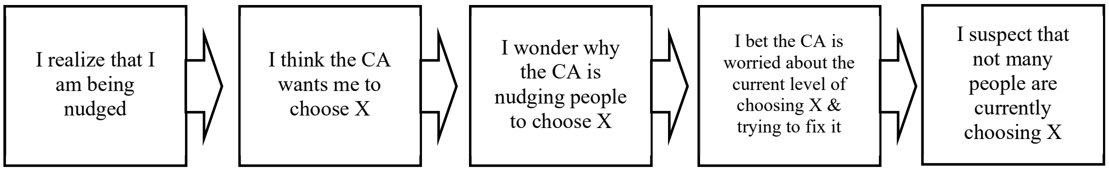

I realize that I am being nudged
I think the CA wants me to choose X
I wonder why the CA is nudging people to choose X
I bet the CA is worried about the current level of choosing X & trying to fix it
I suspect that not many people are currently choosing X

Figure 3-5. *Conceptual model of the series of thoughts that occur to people along the inferential chain explaining the results observed in the present project. Note: "CA" stands for "Choice Architect.*

Ironically, even norms nudges, which operate via social proof, can have this dampening effect on the perceived descriptive social norm.

The present paper documents an effect only on judgment, not on behavior. However, its findings suggest that whenever a nudge triggers negative inferences about the PDSN, a "negative social proof" nudge is thereby also instantiated. The result of this would be to counteract the intended effect of the intervention. This raises the possibility that this phenomenon could be a contributing factor whenever a nudge fails to have its expected result when implemented in the field. Demonstrating this remains the work for future research. Additionally, it is likely that not all nudges and not all situations trigger social sensemaking in the first place. As but one obvious example: If a decision maker is not aware of a nudge (e.g., a friction having been removed from the prior process of choosing the best option), there likely will be no negative effect on the PDSN. Future work should explore the features of nudges that trigger social sensemaking and the boundary conditions of the present phenomenon.

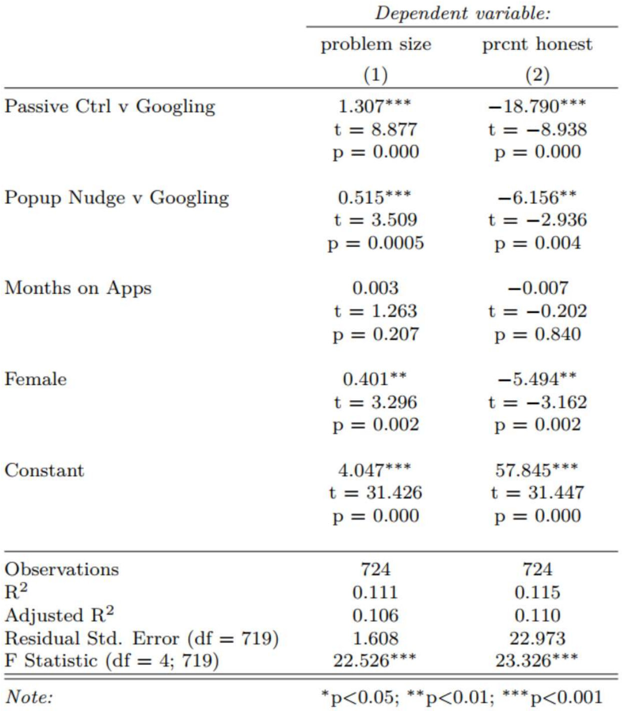
As noted, decades of research have documented the potency of social proof—When people are unsure of what to do, they often assimilate into what is normal. Despite this, several tests of norms nudges have failed or even backfired when implemented in the field (Ashraf et al., 2014; Barankay, 2012; Beshears et al., 2015; Bursztyn & Jensen, 2015; Hennig-Schmidt et al., 2010). Perhaps one way to reconcile these two facts is to note the distinction between norms (the social *information*) and norms nudges (the act of *providing* this information). As we have shown, whatever the effect of the information, sometimes using this information as a nudge can strike people as the choice architect protesting just a little too much.

Appendix

A1. Regression Results from Studies 1 – 3, per Pre-registration Plans

Study 1

Table A3-1: Main Analysis



	Dependent variable:	
	problem size	prcnt honest
	(1)	(2)

Passive Ctrl v Googling	1.307***
	t = 8.877
	p = 0.000
		−18.790***
		t = −8.938
		p = 0.000

Popup Nudge v Googling	0.515***
	t = 3.509
	p = 0.0005
		−6.156**
		t = −2.936
		p = 0.004

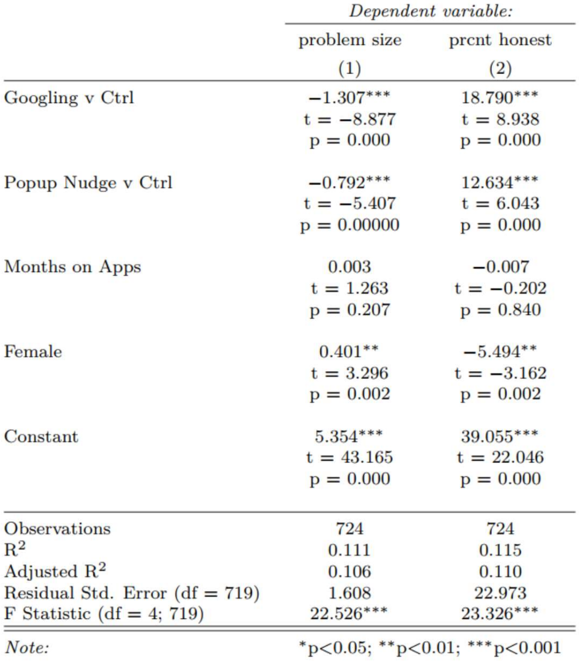
Months on Apps	0.003
	t = 1.263
	p = 0.207
		−0.007
		t = −0.202
		p = 0.840

Female	0.401**
	t = 3.296
	p = 0.002
		−5.494**
		t = −3.162
		p = 0.002

Constant	4.047***
	t = 31.426
	p = 0.000
		57.845***
		t = 31.447
		p = 0.000

Observations	724	724
R^2	0.111	0.115
Adjusted R^2	0.106	0.110
Residual Std. Error (df = 719)	1.608	22.973
F Statistic (df = 4; 719)	22.526***	23.326***
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table A3-2: Secondary Analyses versus Passive Control



	Dependent variable:	
	problem size	prcnt honest
	(1)	(2)

Googling v Ctrl	−1.307***
	t = −8.877
	p = 0.000
		18.790***
		t = 8.938
		p = 0.000

Popup Nudge v Ctrl	−0.792***
	t = −5.407
	p = 0.00000
		12.634***
		t = 6.043
		p = 0.000

Months on Apps	0.003
	t = 1.263
	p = 0.207
		−0.007
		t = −0.202
		p = 0.840

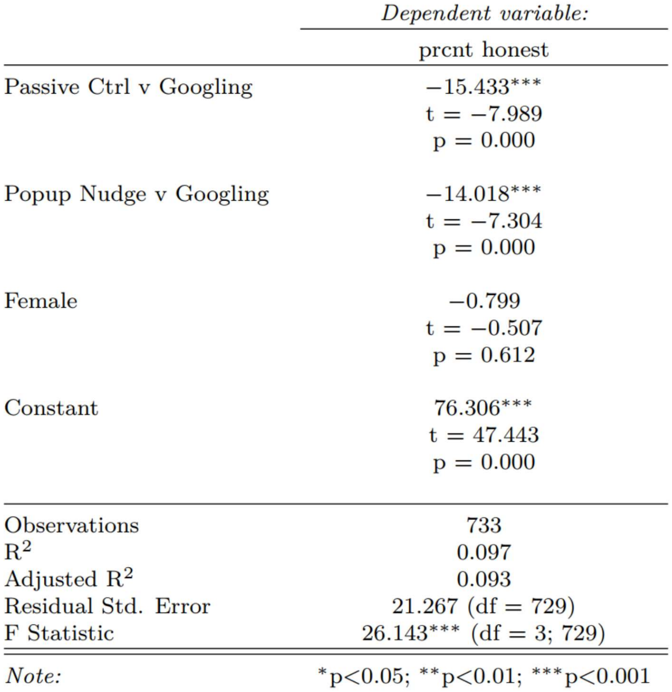
Female	0.401**
	t = 3.296
	p = 0.002
		−5.494**
		t = −3.162
		p = 0.002

Constant	5.354***
	t = 43.165
	p = 0.000
		39.055***
		t = 22.046
		p = 0.000

Observations	724	724
R^2	0.111	0.115
Adjusted R^2	0.106	0.110
Residual Std. Error (df = 719)	1.608	22.973
F Statistic (df = 4; 719)	22.526***	23.326***
Note:	*p<0.05; **p<0.01; ***p<0.001	

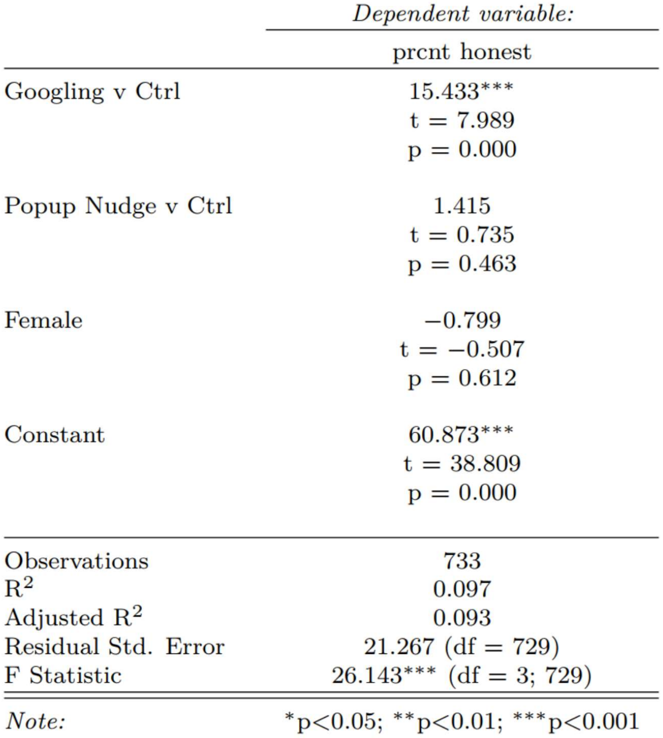
Study 2

Table A3-3: Main Analysis



	Dependent variable:
	prcnt honest
Passive Ctrl v Googling	-15.433^{***}
	$t = -7.989$
	$p = 0.000$
Popup Nudge v Googling	-14.018^{***}
	$t = -7.304$
	$p = 0.000$
Female	-0.799
	$t = -0.507$
	$p = 0.612$
Constant	76.306^{***}
	$t = 47.443$
	$p = 0.000$
Observations	733
R^2	0.097
Adjusted R^2	0.093
Residual Std. Error	21.267 (df = 729)
F Statistic	26.143^{***} (df = 3; 729)
Note:	$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

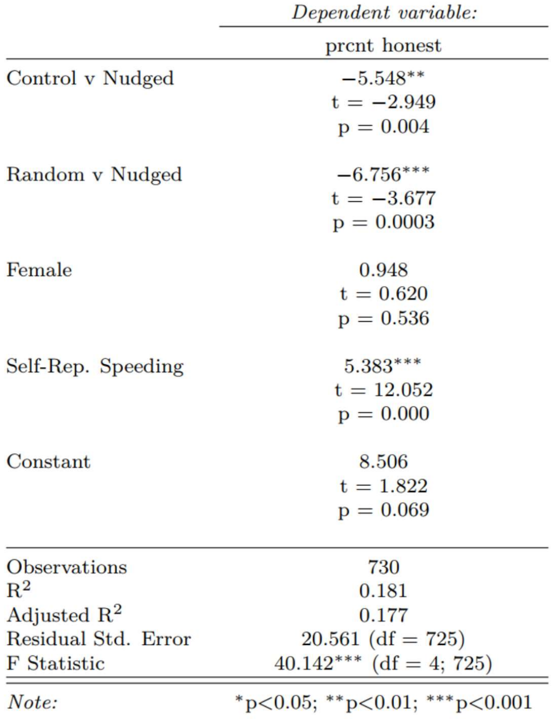
Table A3-4: Secondary Analyses versus Passive Control



	Dependent variable:
	prcnt honest
Googling v Ctrl	15.433^{***}
	$t = 7.989$
	$p = 0.000$
Popup Nudge v Ctrl	1.415
	$t = 0.735$
	$p = 0.463$
Female	-0.799
	$t = -0.507$
	$p = 0.612$
Constant	60.873^{***}
	$t = 38.809$
	$p = 0.000$
Observations	733
R^2	0.097
Adjusted R^2	0.093
Residual Std. Error	21.267 (df = 729)
F Statistic	26.143^{***} (df = 3; 729)
Note:	$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Study 3

Table A3-5: Compared to Nudging



	Dependent variable:
	prcnt honest

Control v Nudged	−5.548**
	t = −2.949
	p = 0.004

Random v Nudged	−6.756***
	t = −3.677
	p = 0.0003

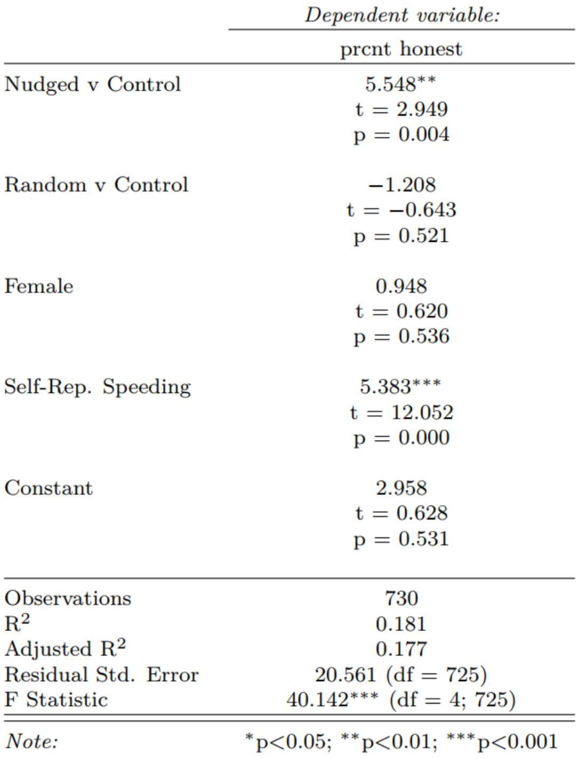
Female	0.948
	t = 0.620
	p = 0.536

Self-Rep. Speeding	5.383***
	t = 12.052
	p = 0.000

Constant	8.506
	t = 1.822
	p = 0.069

Observations	730
R^2	0.181
Adjusted R^2	0.177
Residual Std. Error	20.561 (df = 725)
F Statistic	40.142*** (df = 4; 725)
Note:	*p<0.05; **p<0.01; ***p<0.001

Table A3-6: Compared to Baseline



	Dependent variable:
	prcnt honest

Nudged v Control	5.548**
	t = 2.949
	p = 0.004

Random v Control	−1.208
	t = −0.643
	p = 0.521

Female	0.948
	t = 0.620
	p = 0.536

Self-Rep. Speeding	5.383***
	t = 12.052
	p = 0.000

Constant	2.958
	t = 0.628
	p = 0.531

Observations	730
R^2	0.181
Adjusted R^2	0.177
Residual Std. Error	20.561 (df = 725)
F Statistic	40.142*** (df = 4; 725)
Note:	*p<0.05; **p<0.01; ***p<0.001

REFERENCES

Alavi, S., Habel, J., Schmitz, C., Richter, B., & Wieseke, J. (2018). The risky side of inspirational appeals in personal selling: When do customers infer ulterior salesperson motives? *Journal of Personal Selling & Sales Management*, *38*(3), 323–343. https://doi.org/10.1080/08853134.2018.1447385

Alicke, M. D., LoSchiavo, F. M., Zerbst, J., & Zhang, S. (1997). The person who out performs me is a genius: Maintaining perceived competence in upward social comparison. *Journal of Personality and Social Psychology*, *73*(4), 781–789. https://doi.org/10.1037/0022-3514.73.4.781

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9–10), 1082–1095. https://doi.org/10.1016/j.jpubeco.2011.03.003

Allcott, H., & Rogers, T. (2014). The Short-Run and Long-Run Effects of Behavioral Interventions: Experiment Evidence From Energy Conservation. *American Economic Review*, *104*(10), 3003–3037. https://doi.org/10.1257/aer.104.10.3003

American Community Survey Office. (2013). *American community survey [Data file]*. http://www.census.gov/acs/www/data_documentation/data_main/

Arkes, H., Dawes, R., & Christensen, C. (1986). Factors Influencing the Use of a Decision Rule in a Probabilistic Task. *Organizational Behavior and Human Decision Processes*, *37*, 93–110. https://doi.org/10.1016/0749-5978(86)90046-4

Asensio, O. I., & Delmas, M. A. (2015). Nonprice incentives and energy conservation. *Proceedings of the National Academy of Sciences*, *112*(6), E510–E515. https://doi.org/10.1073/pnas.1401880112

Ashraf, N., Bandiera, O., & Lee, S. S. (2014). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior & Organization*, *100*, 44–63. https://doi.org/10.1016/j.jebo.2014.01.001

Barankay, I. (2012). *Rank Incentives: Evidence from a Randomized Workplace Experiment*. University of Pennsylvania Scholarly Commons. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1074&context=bepp_papers

Bar-Eli, M., Azar, O. H., Ritov, I., Keidar-Levin, Y., & Schein, G. (2007). Action bias among elite soccer goalkeepers: The case of penalty kicks. *Journal of Economic Psychology*, *28*(5), 606–621. https://doi.org/10.1016/j.joep.2006.12.001

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529. https://doi.org/10.1037/0033-2909.117.3.497

Behavioural Insights Team, U. C. O. (2012). *Applying behavioural insights to reduce fraud, error and debt*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/60539/BIT_FraudErrorDebt_accessible.pdf

Bell, D. R., & Bucklin, R. E. (1999). The Role of Internal Reference Points in the Category Purchase Decision. *Journal of Consumer Research*, *26*(2), 128–143. https://doi.org/10.1086/209555

Benabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies*, *70*(3), 489–520.

Berger, J., & Pope, D. (2011). Can losing lead to winning? *Management Science*, *57*(5), 817–827. https://doi.org/10.1287/mnsc.1110.1328

Berkowitz, L. (1972). Social norms, feelings, and other factors affecting helping and altruism. *Advances in Experimental Social Psychology*, *6*(C), 63–108. https://doi.org/10.1016/S0065-2601(08)60025-8

Beshears, J., Choi, J. J., Laibson, D., Madrian, B. C., & Milkman, K. L. (2015). The Effect of Providing Peer Information on Retirement Savings Decisions. *Journal of Finance*, *70*(3), 1161–1201. https://doi.org/10.1111/jofi.12258

Bicchieri, C., & Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public Choice*. https://doi.org/10.1007/s11127-019-00684-6

Bogard, J. E., Delmas, M. A., Goldstein, N. J., & Vezich, I. S. (2020). Target, distance, and valence: Unpacking the effects of normative feedback. *Organizational Behavior and Human Decision Processes*, *161*, 61–73. https://doi.org/10.1016/j.obhdp.2020.10.003

Bollinger, B., & Gillingham, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science*, *31*(6), 900–912. https://doi.org/10.1287/mksc.1120.0727

Brent, D. A., Cook, J. H., & Olsen, S. (2015). Social Comparisons, Household Water Use, and Participation in Utility Conservation Programs: Evidence from Three Randomized Trials. *Journal of the Association of Environmental and Resource Economists*, *2*(4), 597–627. https://doi.org/10.1086/683427

Brown, C. L., & Krishna, A. (2004). The Skeptical Shopper: A Metacognitive Account for the Effects of Default Options on Choice. *Journal of Consumer Research*, *31*.

Burson, K. A. (2007). Consumer-product skill matching: The effects of difficulty on relative self-assessment and choice. *Journal of Consumer Research*, *34*(1), 104–110. https://doi.org/10.1086/513051

Bursztyn, L., & Jensen, R. (2015). How Does Peer Pressure Affect Educational Investments? *The Quarterly Journal of Economics*, *130*(3), 1329–1367.

Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics Using Stata*. Stata Press.

Campbell, M. C., & Kirmani, A. (2000). Consumers' Use of Persuasion Knowledge: The Effects of Accessibility and Cognitive Capacity on Perceptions of an Influence Agent. *Journal of Consumer Research*, *27*(1), 69–83. https://doi.org/10.1086/314309

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Centraal Bureau voor de Statistiek. (2017, July 26). *Ontwikkeling donorregistraties 2016* [Webpagina]. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/nieuws/2017/30/ontwikkeling-donorregistraties-2016

Chen, V. L., Delmas, M. A., Kaiser, W. J., & Locke, S. L. (2015). What can we learn from high-frequency appliance-level energy metering? Results from a field experiment. *Energy Policy*, *77*, 164–175. https://doi.org/10.1016/j.enpol.2014.11.021

Cialdini, R. B. (2021). *Influence, New and Expanded: The Psychology of Persuasion* (4th ed.). Harper Business.

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence*, *1*(1), 3–15. https://doi.org/10.1080/15534510500181459

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*(1), 591–621.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and re-evaluation. *Advances in Experimental Social Psychology*, *24*, 201–234. https://doi.org/10.1016/ S0065-2601(08)60330-5

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity, and compliance. In D.T. Gilbert, S.T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (pp. 151–192). McGraw-Hill.

Cryder, C. E., John London, A., Volpp, K. G., & Loewenstein, G. (2010). Informative inducement: Study payment as a signal of risk. *Social Science & Medicine*, *70*(3), 455–464. https://doi.org/10.1016/j.socscimed.2009.10.047

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical Versus Actuarial Judgment. *Science*, *243*(4899), 1668–1674. JSTOR.

Delmas, M. A., Fischlein, M., & Asensio, O. I. (2013). Information strategies and energy conservation behavior: A meta-analysis of experimental studies from 1975 to 2012. *Energy Policy*, *61*, 729–739. https://doi.org/10.1016/j.enpol.2013.05.109

Delmas, M. A., & Lessem, N. (2014). Saving power to conserve your reputation? The effectiveness of private versus public information. *Journal of Environmental Economics and Management*, *67*(3), 353–370. https://doi.org/10.1016/j.jeem.2013.12.009

Demetis, D., & Lee, A. (2018). When humans using the IT artifact becomes IT using the human artifact. *Journal of the Association for Information Systems*, *19*(10), 929–952. https://doi.org/10.17705/1jais.00513

Deserranno, E. (2019). Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda. *American Economic Journal: Applied Economics*, *11*(1), 277–317.

Dietvorst, B., & Bharti, S. (2020). People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error. *Psychological Science*, 0956797620948841. https://doi.org/10.1177/0956797620948841

Dietvorst, B., Simmons, J., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B., Simmons, J. P., & Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dinner, I., Johnson, E. J., Goldstein, D. G., & Liu, K. (2011). Partitioning default effects: Why people choose not to choose. *Journal of Experimental Psychology: Applied*, *17*(4), 332.

Everett, J. A. C., Caviola, L., Kahane, G., Savulescu, J., & Faber, N. S. (2015). Doing good by doing nothing? The role of social norms in explaining default effects in altruistic contexts. *European Journal of Social Psychology*, *45*(2), 230–241. https://doi.org/10.1002/ejsp.2080

Fehr, E., & List, J. A. (2004). The Hidden Costs and Returns of Incentives—Trust and Trustworthiness Among Ceos. *Journal of the European Economic Association*, *2*(5), 743–771.

Feldman, G., & Albarracín, D. (2017). Norm theory and the action-effect: The role of social norms in regret following action and inaction. *Journal of Experimental Social Psychology*, *69*, 111–120. https://doi.org/10.1016/j.jesp.2016.07.009

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Frey, B. S., & Oberholzer, F. (1997). The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out. *The American Economic Review*, *87*(4), 746–755.

Friestad, M., & Wright, P. (1994). The Persuasion Knowledge Model: How People Cope with Persuasion Attempts. *Journal of Consumer Research*, *21*(1), 1–31.

Gates, S. W., Perry, V. G., & Zorn, P. M. (2002). Automated underwriting in mortgage lending: Good news for the underserved? *Housing Policy Debate*, *13*(2), 369–391. https://doi.org/10.1080/10511482.2002.9521447

Gilbert, D. T., Krull, D. S., & Pelham, B. W. (1988). Of Thoughts Unspoken Social Inference and the Self-Regulation of Behavior. *Journal of Personality and Social Psychology*, *55*(5), 685–694.

Gillingham, K., & Palmery, K. (2014). Bridging the energy efficiency gap: Policy insights from economic theory and empirical evidence. *Review of Environmental Economics and Policy*, *8*(1), 18–38. https://doi.org/10.1093/reep/ret021

Göckeritz, S., Schultz, P. W., Rendón, T., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2010). Descriptive normative beliefs and conservation behavior: The moderating roles of personal involvement and injunctive normative beliefs. *European Journal of Social Psychology*, *40*(3), 514–523. https://doi.org/10.1002/ejsp.643

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*(3), 472–482. https://doi.org/10.1086/586910

Grice, H. P. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Speech Acts* (pp. 41–58). Brill.

Grønhøj, A., & Thøgersen, J. (2011). Feedback on household electricity consumption: Learning and social influence processes. *International Journal of Consumer Studies*, *35*(2), 138–145. https://doi.org/10.1111/j.1470-6431.2010.00967.x

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as Reference Points. *Cognitive Psychology*, *38*, 79–109.

Hennig-Schmidt, H., Rockenbach, B., & Sadrieh, A. (2010). In Search of Workers' Real Effort Reciprocity – A Field and a Laboratory Experiment. *Journal of the European Economic Association*, *8*(4), 817–837.

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, *3*(02), 159–186. https://doi.org/10.1017/bpp.2018.43

Jachimowicz, J. M., Hauser, O. P., O'Brien, J. D., Sherman, E., & Galinsky, A. D. (2018). The critical role of second-order normative beliefs in predicting energy conservation. *Nature Human Behaviour*, *2*(10), 757–764. https://doi.org/10.1038/s41562-018-0434-0

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153. https://doi.org/10.1037/0033-295X.93.2.136

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under Uncertainty* (1st ed., pp. 201–208). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.015

Kamenica, E. (2008). Contextual Inference in Markets: On the Informational Content of Product Lines. *The American Economic Review*, *98*(5), 2127–2149. JSTOR.

Karlin, B., Zinger, J. F., & Ford, R. (2015). The effects of feedback on energy conservation: A meta-analysis. *Psychological Bulletin*, *141*(6), 1205–1227. https://doi.org/10.1037/a0039650

Kerr, S., & Landauer, S. (2004). Using stretch goals to promote organizational effectiveness and personal growth: General Electric and Goldman Sachs. In *Academy of Management Executive* (Vol. 18, Issue 4, pp. 134–138). https://doi.org/10.5465/AME.2004.15268739

Kivetz, R., Urminsky, O., & Zheng, Y. (2006). The Goal-Gradient Hypothesis Resurrected: Purchase Acceleration, Illusionary Goal Progress, and Customer Retention. *Journal of Marketing Research*, *43*(1), 39–58. https://doi.org/10.1509/jmkr.43.1.39

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Krijnen, J. M. T., Tannenbaum, D., & Fox, C. R. (2017). Choice architecture 2.0: Behavioral policy as an implicit social interaction. *Behavioral Science & Policy*, *3*(2), i–18. https://doi.org/10.1353/bsp.2017.0010

Larrick, R., & Boles, T. (1995). Avoiding regret in decisions with feedback: A negotiation example. *Organizational Behavior and Human Decision Processes*, *63*(1), 87–97.

Lee, C. Y. (2003). Firm density and industry R&D intensity: Theory and evidence. *Review of Industrial Organization*, *22*(2), 139–158. https://doi.org/10.1023/A:1022965830769

Lewis, M. A., & Neighbors, C. (2006). Who is the typical college student? Implications for personalized normative feedback interventions. *Addictive Behaviors*, *31*(11), 2120–2126. https://doi.org/10.1016/j.addbeh.2006.01.011

Liberman, N., & Förster, J. (2008). Expectancy, value and psychological distance: A new look at goal gradients. *Social Cognition*, *26*(5), 515–533. https://doi.org/10.1521/soco.2008.26.5.515

Locke, E. A., & Latham, G. P. (2006). New Directions in Goal-Setting Theory. *CURRENT DIRECTIONS IN PSYCHOLOGICAL SCIENCE*, *15*(5).

Lockwood, P., & Kunda, Z. (1997). *Superstars and Me: Predicting the Impact of Role Models on the Self*. *73*(1), 91–103.

Lockwood, P., & Kunda, Z. (1999). Increasing the salience of one's best selves can undermine inspiration by outstanding role models. *Journal of Personality and Social Psychology*, *76*(2), 214–228. https://doi.org/10.1037/0022-3514.76.2.214

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Maitlis, S., & Christianson, M. (2014). Sensemaking in Organizations: Taking Stock and Moving Forward. *Academy of Management Annals*, *8*(1), 57–125. https://doi.org/10.5465/19416520.2014.873177

Major, B., Testa, M., & Blysma, W. H. (1991). Responses to upward and downward social comparisons: The impact of esteem-relevance and perceived control. In *Social Comparison: Contemporary Theory and Research* (pp. 237–260). Lawrence Erlbaum Associates, Inc.

Manning, A. D., Lindenmayer, D. B., & Fischer, J. (2006). Stretch goals and backcasting: Approaches for overcoming barriers to large-scale ecological restoration. *Restoration Ecology*, *14*(4), 487–492. https://doi.org/10.1111/j.1526-100X.2006.00159.x

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370.

Mckenzie, C. R. M. M., Liersch, M. J., & Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. *Psychological Science*, *17*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2006.01721.x

Meeker, D., Linder, J. A., Fox, C. R., Friedberg, M. W., Persell, S. D., Goldstein, N. J., Knight, T. K., Hay, J. W., & Doctor, J. N. (2016). Effect of behavioral interventions on inappropriate antibiotic prescribing among primary care practices: A randomized clinical trial. *JAMA*, *315*(6), 562. https://doi.org/10.1001/jama.2016.0275

Miller, D. T., & McFarland, C. (1986). Counterfactual Thinking and Victim Compensation: A Test of Norm Theory. *Personality and Social Psychology Bulletin*, *12*(4), 513–519. https://doi.org/10.1177/0146167286124014

Misch, A., Over, H., & Carpenter, M. (2016). I Won't Tell: Young Children Show Loyalty to Their Group by Keeping Group Secrets. *Journal of Experimental Child Psychology*, *142*, 96–106. https://doi.org/10.1016/j.jecp.2015.09.016

Mortensen, C. R., Neel, R., Cialdini, R. B., Jaeger, C. M., Jacobson, R. P., & Ringel, M. M. (2019). Trending Norms: A Lever for Encouraging Behaviors Performed by the Minority.

Social Psychological and Personality Science, *10*(2), 201–210. https://doi.org/10.1177/1948550617734615

Navarro-Martinez, D., Salisbury, L. C., Lemon, K. N., Stewart, N., Matthews, W. J., & Harris, A. J. L. (2011). Minimum Required Payment and Supplemental Information Disclosure Effects on Consumer Debt Repayment Decisions. *Journal of Marketing Research*, *48*(SPL), S60–S77. https://doi.org/10.1509/jmkr.48.SPL.S60

Neighbors, C., LaBrie, J. W., Hummer, J. F., Lewis, M. A., Lee, C. M., Desai, S., Kilmer, J. R., & Larimer, M. E. (2010). Group identification as a moderator of the relationship between perceived social norms and alcohol consumption. *Psychology of Addictive Behaviors*, *24*(3), 522–528. https://doi.org/10.1037/a0019944

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin*, *34*(7), 913–923. https://doi.org/10.1177/0146167208316691

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Orhun, A. Y. (2018). Perceived motives and reciprocity. *Games and Economic Behavior*, *109*, 436–451. https://doi.org/10.1016/j.geb.2018.01.002

Panko, R. (2018). *The Popularity of Google Maps: Trends in Navigation Apps in 2018*. The Manifest. https://themanifest.com/mobile-apps/popularity-google-maps-trends-navigation-apps-2018

Rainie, L., & Anderson, J. (2017). Code-Dependent: Pros and Cons of the Algorithm Age. *Pew Research Center*. http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age

Ritov, I. (1996). Probability of Regret: Anticipation of Uncertainty Resolution in Choice. *Organizational Behavior and Human Decision Processes*, *66*(2), 228–236. https://doi.org/10.1006/obhd.1996.0051

Ritov, I., & Baron, J. (1995). Outcome Knowledge, Regret, and Omission Bias. *Organizational Behavior and Human Decision Processes*, *64*(2), 119–127.

Rogers, T., & Feller, A. (2016). Discouraged by Peer Excellence: Exposure to Exemplary Peer Performance Causes Quitting. *Psychological Science*, *27*(3), 365–374. https://doi.org/10.1177/0956797615623770

Rosenfeld, M. J., Thomas, R. J., & Hausen, S. (2019). Disintermediating your friends: How online dating in the United States displaces other ways of meeting. *Proceedings of the National Academy of Sciences*, *116*(36), 17753–17758. https://doi.org/10.1073/pnas.1908630116

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, *32*(5), 880–892. https://doi.org/10.1037/0022-3514.32.5.880

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science*, *12*(3), 185–190.

Rottenstreich, Y., & Shu, S. B. (2004). The Connections Between Affect nad Decision Making: Nine Resulting Phenomena. In *Blackwell Handbook of Judgment and Decision Making* (pp. 444–463). Blackwell Publishing Ltd.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Schachter, S. (1959). *The psychology of affiliation: Experimental studies of the sources of gregariousness* (p. 141). Stanford Univer. Press.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 1–19. https://doi.org/10.1080/12460125.2020.1819094

Schultz, P. W. (1999). Changing behavior with normative feedback interventions: A field experiment on curbside recycling. *Basic and Applied Social Psychology*, *21*(1), 25–36. https://doi.org/10.1207/s15324834basp2101_3

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychological Science*, *18*(5), 429–434. https://doi.org/10.1111/j.1467-9280.2007.01917.x

Schwarz, N. (1994). Judgment in a Social Context: Biases, Shortcomings, and the Logic of Conversation. In *Advances in Experimental Social Psychology* (Vol. 26, pp. 123–162). Elsevier. https://doi.org/10.1016/S0065-2601(08)60153-7

Shang, J., & Croson, R. (2009). A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods. *The Economic Journal*, *119*(540), 1422–1439. https://doi.org/10.1111/j.1468-0297.2009.02267.x

Sherif, M. (1936). The psychology of social norms. In *The psychology of social norms.* Harper.

Simonson, I. (1992). The Influence of Anticipating Regret and Responsibility on Purchase Decisions. *Journal of Consumer Research*, *19*(1), 105–118.

Tankard, M. E., & Paluck, E. L. (2016). Norm Perception as a Vehicle for Social Change: Vehicle for Social Change. *Social Issues and Policy Review*, *10*(1), 181–211. https://doi.org/10.1111/sipr.12022

Tannenbaum, D., Fox, C. R., & Goldstein, N. J. (2017). *Partitioning Menu Items to Nudge Single-item Choice*. github.io. https://davetannenbaum.github.io/documents/pdepend.pdf

Tesser, A. (1988). Toward a Self-Evaluation Maintenance Model of Social Behavior. *Advances in Experimental Social Psychology*, *21*(C), 181–227. https://doi.org/10.1016/S0065-2601(08)60227-0

Tesser, A. (1991). Emotion in social comparison and reflection processes. In J. Suls & T. A. Wills (Eds.), *Social comparison: Contemporary theory and research.* (pp. 115–145).

Tesser, A., & Campbell, J. (1983). Self-definition and self-evaluation maintenance. *Psychological Perspectives on the Self*, 1–31.

Tybout, A. M., & Yalch, R. F. (1980). The effect of experience: A matter of salience? *Journal of Consumer Research*, *6*(4), 406. https://doi.org/10.1086/208783

Verlegh, P. W. J., Ryu, G., Tuk, M. A., & Feick, L. (2013). Receiver responses to rewarded referrals: The motive inferences framework. *Journal of the Academy of Marketing Science*, *41*(6), 669–682. https://doi.org/10.1007/s11747-013-0327-8

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Weick, K. E. (1995). *Sensemaking in Organizations*. Sage.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the Process of Sensemaking. *Organization Science*, *16*(4), 409–421. https://doi.org/10.1287/orsc.1050.0133

Wernerfelt, B. (1995). A Rational Reconstruction of the Compromise Effect: Using Market Data to Infer Utilities. *Journal of Consumer Research*, *21*(4), 627–633. JSTOR.

Wright, P. (2002). Marketplace Metacognition and Social Intelligence. *Journal of Consumer Research*, *28*(4), 677–682. https://doi.org/10.1086/338210

Zeelenberg, M. (1999). Anticipated regret, expected feedback and behavioral decision making. *Journal of Behavioral Decision Making*, *12*(2), 93–106. https://doi.org/10.1002/(SICI)1099-0771(199906)12:2<93::AID-BDM311>3.0.CO;2-S

Zeelenberg, M., & Pieters, R. (2004). Consequences of regret aversion in real life: The case of the Dutch postcode lottery. *Organizational Behavior and Human Decision Processes*, *93*(2), 155–168. https://doi.org/10.1016/j.obhdp.2003.10.001

Zeelenberg, M., van Dijk, W. W., van der Pligt, J., Manstead, A. S. R., van Empelen, P., & Reinderman, D. (1998). Emotional Reactions to the Outcomes of Decisions: The Role of Counterfactual Thought in the Experience of Regret and Disappointment. *Organizational Behavior and Human Decision Processes*, *75*(2), 117–141. https://doi.org/10.1006/obhd.1998.2784