Rhetoric and Social Norms

by

Daphne Chang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Erin Krupka, Chair
Associate Professor Stephen Leider
Assistant Professor Yesim Orhun
Associate Professor Tanya Rosenblat

Daphne Chang

daphnec@umich.edu

ORCID iD: 0000-0003-2599-2598

For my family and MG

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Erin Krupka for her guidance and support throughout my study as well as the dissertation writing process. I am also extremely grateful to my committee, Stephen Leider, Yesim Orhun, and Tanya Rosenblat, for their invaluable feedback and insightful advice. Many thanks to Ryan Burton, Sam Carton, Linfeng Li, Chanda Phelan, and Hariharan Subramonyam for their helpful contributions. This dissertation also benefited greatly from discussions with members of the Behavioral and Experimental Economics lab at the University of Michigan. I also thank my family for their encouragement and understanding. Lastly, I'm deeply indebted to Michael Gelman. This dissertation would not have been possible without his unwavering patience and support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Social norms characterize what is expected of an individual, while rhetoric employs language to communicate those norms and to shape beliefs (*Kahneman*, 2000). Social norms and rhetoric have been widely used by public and private institutions to encourage desirable behavior, such as lowering energy consumption (e.g. *Schultz et al.*, 2007; *Nolan et al.*, 2008) or increasing voter turn-out (e.g. *Gerber et al.*, 2008; *Gerber and Rogers*, 2009).

However, these strategic uses of norms and rhetoric can have detrimental consequences, particularly when they are deployed to create divisions in society or to bias beliefs about the norms. This dissertation examines the purposeful and strategic use of social norms and rhetoric to influence behavior. Specifically, I focus on the degree to which individuals understand and choose to use these influences to sway the behavior of others. I explore this in the political context, where these influences can be used to create divisions in society along party lines. I also examine this in the private information disclosure context, where these influences can be leveraged to create biased beliefs about the disclosure norms.

The first chapter is motivated by the observation that public support can wax and wane depending on how a situation is described. For example, public support can vary by whether tax policy reform is described as tax-relief or a tax-cut. Understanding how frames influence choice can provide a way to make the consequences of framing more predictable and aid in ameliorating our susceptibility to them. I propose that this rhetorical technique of "framing" influences the decision-maker's choice by

invoking different perceptions of socially appropriate behavior, or norms. I examine this in a political context and test how the social norms evoked by politically charged language (the "tax frame") or non-politically charged language (the "neutral frame") can cause behavior to change differently depending on a person's political identity. I find that the frames have significant impact on both the norms and choices of Democratic and Republican participants. Tax-framed Democratic norms favor more equal allocations relative to neutrally-framed Democratic norms. In contrast, tax-framed Republican norms favor keeping allocations at the status quo while neutrally-framed Republican norms do not. But differences between tax-framed Democratic and Republican norms disappear when the frame is neutral. The behavior of tax- framed (neutrally-framed) Democrats and Republicans closely follow their proscribed tax-framed (neutrally-framed) norms. This paper demonstrates how political rhetoric may be used to create division in society among individuals who would have otherwise made the same choice. The strategic choice of language influences the social norms that is communicated by it, which then affects behavior. This is joint work with Roy Chen and Erin Krupka.

The second chapter examines how the strategic use of social norms in online platform designs can shape beliefs about information disclosure and influence disclosure behavior. Nudges that strategically leverage social norms are often incorporated in a website's design for the benefit of the users. But these nudges may also be used for malevolent means, encouraging users to disclose more information than they are aware that they are sharing or would otherwise decline to share. However, the mechanisms through which these interface designs affect user disclosure are still poorly understood. This paper empirically tests how changes in the most common behavior observed by an user affects his or her personal beliefs about acceptable behavior, which then influence his or her choice. Using an experiment, I nudge participants' personal beliefs by manipulating the perceived social norms of a social network site.

I do so by showing them the social network site with either more or less provocative images. I then ask the participants to either share personal information or to advise another user on whether she should share her personal information. I find that, relative to a less provocative set of images, a more provocative set of images nudges an individual's personal views of what is appropriate to share, which in turn increases the probability that the individual divulges personal information and the probability that the individual advise others to do the same. This paper establishes the causal pathway between a social nudge that leverages social norms on an individual's disclosure behavior. This paper further adds insight into how newcomers in a community might adopt the existing norms. This is joint work with Eytan Adar, Erin Krupka, and Alessandro Acquisti.

The third chapter examines whether everyday individuals can use rhetoric effectively to influence beliefs and decision outcomes. A promotion decision is complex and managers must anticipate how candidates will perform in a familiar, but novel task. Managers frequently rely on each candidate's current performance to inform their decisions. However, the measures of a candidate's current performance are often imperfect signals of the candidate's actual effort and ability. Candidates may also influence managers' hiring choice by responding with statements about themselves or their competitors. In this paper, I test whether the ways in which candidates for promotion respond to information about their past performance can influence managers' decisions. I further test whether candidates can strategically identify and use the most effective response to influence managers' decisions. Using an experiment, I vary the information that a manager receives about the relative performance of two candidates on an earlier task. The manager then receives costless, unverifiable responses (cheap talk responses) from the two candidates. These responses may either be self-promoting ("sweet"), other-defaming ("mean") or a neutral statement ("neutral"). I show that such cheap talk affects the manager's beliefs of the candi-

dates' past performance and the manager's hiring choice. Moreover, the influence of these responses is dependent on the contents of the preceding information. Lastly, I show evidence that responses influence the manager's decision via belief as well as non-belief channels.

CHAPTER I

Rhetoric Matters: A Social Norms Explanation for the Anomaly of Framing

1.1 Introduction

Framing matters. Though most people think that we have stable preferences for how we approach choices, there is ample evidence that certain words or ways of phrasing things can cause us to change our preferences. The rhetorical technique of "framing" is defined as the act of describing a situation in such a way as to change the decision-maker's conception of the acts, outcomes and associated contingencies for that situation (*Tversky and Kahneman*, 1981).[1] Previous research has documented the existence of framing in a number of contexts. For example, *Tversky and Kahneman* (1981); *Larrick and Blount* (1997), and *Dufwenberg et al.* (2011a) provide evidence that two versions of a decision problem that are transparently equivalent evoke different preferences when considered separately. *Rugg* (1941) demonstrates

[1] As noted by *Kahneman* (2000), there are several different ways to interpret "framing effect," including an experimental manipulation that changes the description of the situation and a characterization of how players in a game conceptualize strategies. In our experimental design, we adopt the former interpretation (see also *Dufwenberg et al.*, 2011a). Our design uses what *Larrick and Blount* (1997) define as "procedural framing," where actions are described in different ways for structurally equivalent allocation procedures. As an example, in *Liberman et al.* (2004) the same prisoner's dilemma game is framed as a "Wall Street Game" and a "Community Game." This difference in framing leads to a difference in participants' choices. See also *Cookson* (2000); *Rege and Telle* (2004); *Dufwenberg et al.* (2011a); *Ellingsen et al.* (2012) and *Banerjee* (2016).

the effectiveness of framing in public opinion polling. In his study, 62% of respondents answered "no" to the question "Do you think the United States should allow public speeches against democracy?", but only 46% of respondents answered "yes" to the question "Do you think the United States should forbid public speeches against democracy?". Similarly, *Nelson et al.* (1997a) find that whether a rally by the Ku Klux Klan is framed as a free speech issue or a disruption of public order affects respondents' tolerance levels for the Klan.

Because frames impact choice, our understanding of the mechanisms by which they do so can provide a way to make the consequences of framing more predictable. One explanation is that framing effects are driven by the asymmetry in how different information is encoded and processed (*Tversky and Kahneman*, 1981). An alternative explanation is that frames activate existing information in an individual's memory, and subsequently influence how that individual weighs her beliefs (*Nelson et al.*, 1997b). In this paper, we propose an additional explanation: frames invite different interpretations of acts and outcomes because they evoke different norms.

The social identity model provides a window through which to observe a mechanism for the effect of framing (*Akerlof and Kranton*, 2000, 2005). Social identity describes the part of an individual's sense of self that stems from their perceived membership with a social group. The utility derived from social identity comes from a desire to comply with the norms for an individual's social identity (*Akerlof and Kranton*, 2000).[2]

[2]Each social group has a set of corresponding normative prescriptions (norms) for behavior that characterize how members of that group ought to behave in a particular situation. Social identity-dependent choice can explain a host of observed social phenomena such as ingroup bias (*Terry and O'Brien*, 2001; *Wichardt*, 2008; *Goette et al.*, 2012), persistence of stereotypes (*Steele and Aronson*, 1995; *Shih et al.*, 1999, 2006; *Afridi et al.*, 2015), and labor disputes (*Akerlof and Kranton*, 2005). In addition, it has been shown to affect cooperation (*Eckel and Grossman*, 2005; *Goette et al.*, 2006; *Charness et al.*, 2007), coordination (*Weber*, 2006; *Chen and Chen*, 2011; *McCarter and Sheremeta*, 2013; *Chen et al.*, 2014), behavior in markets (*Li et al.*, 2011; *Gneezy et al.*, 2012), and voting (*Pickup et al.*, 2016, 2018a,b). Both field and laboratory experiments show that inducing a social identity or making an existing identity salient can shift time, risk and other-regarding preferences (*Chen and Li*, 2009; *Benjamin et al.*, 2010; *Butler*, 2014).

In our experiment, we compare subject responses in a series of dictator games for those given a tax frame with those given a neutral frame. The difference in framing allows us to make a subject's social identity salient and to evoke the associated norms for that identity. We then collect data in a separate treatment to elicit identity-dependent norms. While we follow the work of *Krupka and Weber* (2013a), our primary focus is on the impact of a tax frame on norms for the dictator games.

We show that these frames cause respondents to apply different norms to the situation and cause them to act differently. We document this effect in the context of U.S. political identity (Republicans and Democrats). We then test whether a social identity model can explain our results. Two tests provide evidence that a social identity model predicts behavior better than a benchmarking model without norms.

Our main contribution is our experimental evidence on how frames evoke norms. The finding offers an additional mechanism, frame evoked norms, by which to predict how unstable preferences will be impacted by a frame. In addition, a novel application of the norm elicitation method developed by *Krupka and Weber* (2013a), allows for sharper predictions regarding the likely impact of frames on behavior. In application to politics, this result raises the interesting question of how divided we really are?[3] The evidence suggests that Democrats and Republicans have different views on redistribution but that these differences seem to disappear when political identities are not made salient. It follows then that a key activity of political parties is to use rhetoric to frame choices for their members and pursue identity politics. These results can significantly advance the study of the post-neoclassical anomaly of "apparently" unstable preferences. It also furthers the study of rhetoric on behavior and political

[3]One result that suggests that the impact of the frame depends on identity comes from *Hardisty et al.* (2010). In the study, the payments for an environmental cost are described as either "earmarked taxes" or "offsets." They find that the framing of the payment changes expressed preferences for it by Republicans and Independents, but not by Democrats. The authors interpret their findings as an indication that the frame-induced behavior changes stem from changes in the norms that subjects apply to the situation. See also *Blount and Larrick* (2000), *Koch* (1998) and *Allison et al.* (1996). In other words, one reason why frames invite different interpretations of acts and outcomes is because they evoke different norms.

discourse.

Our second contribution is to advance how we can study social identity by eliciting identity-dependent norms. In our experiments, we use the frame treatment to evoke identity-dependent norms. In our experimental design, we rely on the same causality argument proposed by *Krupka and Weber* (2013a): changes in norms predict changes in behavior in otherwise identical dictator games. However, unlike *Krupka and Weber* (2013a), we use a framing treatment to evoke *identity-dependent* norms in order to show that these identity-dependent norms cause behavior changes that are consistent with the social identities. This novel approach introduces a new way to study a broad range of questions relating to the impact of social identity on behavior.

1.2 Theoretical Framework

The social identity model provides a theoretical framework to elucidate one mechanism, norms, through which frames can affect choice (*Akerlof and Kranton*, 2000, 2005). In their study, *Akerlof and Kranton* (2005) note that "...much of utility depends not only on what economists normally think of as *tastes*, but also on *norms* as to how people think that they and others *should* behave... views as to how people should behave depends upon the particular *situation*...". Moreover, norms for how one should behave vary with one's social identity. In their model, a person's identity is seen in the context of gains and losses in utility that result from behavior that conforms to or departs from the norms for that identity in that situation. [4]

This utility is separated into a value placed on monetary payoffs (which are affected only by actions $\mathbf{a} = (a_i, \mathbf{a}_{-i})$) and on adhering to social norms (N). These norms are

[4] *Akerlof and Kranton* (2005) further notes that "The combination of *identity*, *social category*, *norms* and *ideal* allows parsimonious modeling of how utility functions change as people adopt different mental frames of themselves — that is, as they take on different possible identities. Economists have recently adapted from psychology the idea that utility depends upon how a situation is *framed* (Kahneman and Tversky, 1979). Identity describes one special way in which people frame their situation."

affected by an individual's actions (a_i), everyone's social identities ($\mathbf{I} = (I_i, \mathbf{I}_{-i})$), and the situation ($s$):

$$U_i(\mathbf{a}, \mathbf{I}, s) = V_i(a_i|\mathbf{a}_{-i}) + \gamma_i N(a_i|\mathbf{I}, s), \tag{1.1}$$

where V captures a subject's utility over her monetary payoff, and is not dependent on either social identity or the situation.[5]

In the above specification, $N(\cdot)$ is the social norms function that maps utility over the appropriateness of an action in situation s undertaken by individual i (*Krupka and Weber*, 2013a). In other words, when a person's social identity or situation changes, so does that individual's shared view of the appropriateness of the actions. This model assumes that identity-dependent social norms vary at the group level and, furthermore, are both exogenous and given at the individual level.[6]

Finally, the γ_i term reflects the degree to which person i cares about complying with the social norms for her identity. In this model, the degree to which a person cares about adhering to any social norm is fixed. Intuitively, if an individual is characterized as a strong "norm follower," then she will be a strong "norm follower" in any situation.

We follow *Akerlof and Kranton* (2005) in defining a situation as the context of "...when, where, how and between whom a transaction takes place."[7] We posit that

[5]This formalization of the first term in the utility function follows *Akerlof and Kranton* (2005), who write "In a standard economic model, an individual's preferences are fixed, and utility depends only on pecuniary variables." Regarding the second term, though *Akerlof and Kranton* (2000) characterize utility stemming from one's own identity only, we include own and others' identity in the norm function as a way of capturing the idea that the norms of behavior are determined by the identities of all participants. This is alluded to in *Akerlof and Kranton* (2000) in footnote 5: "...individual's self-concept may be formed by seeing oneself through the eyes of others..." This also finds resonance in *Tajfel and Turner* (1979), where they describe how one comes to hold an identity: "...the essential criteria for group membership...are that the individuals concerned define themselves and are defined by others as members of a group." This idea is echoed in more recent papers, such as *Barr et al.* (2018).

[6]The endogenous selection of social identity is sometimes possible, as with choosing one's profession, and sometimes not possible, as with race or gender (cf. *Akerlof and Kranton*, 2000). Endogenous norm formation is not treated here, but we note that norm formation is likely to take some time, and therefore at a particular point in time, it is reasonable to think of the norm as given.

[7]See also *Ellingsen and Mohlin* (2014) who define a situation as a "shared view of the set of participants and the relevant set of actions."

framing can change this situation in at least two unique ways. First, it does so by changing an individual's perception of the associated acts and outcomes. For example, if we alter the framing of the standard dictator game by changing the placement of the initial endowment so that it rests with the non-active second player, then a dictator must take money from this player to achieve a positive payoff for himself. Essentially, a payoff obtained in a dictator game through giving (as in the standard dictator game) is perceived differently from the same payoff obtained through taking (as in the altered dictator game). Second, frames can also change our situation by evoking a social identity and its associated norms. For example, when a dictator game is described as a tax redistribution, it may evoke a person's political identity, making any transfer feel like a "handout." We use both of these changes in our treatment.

1.3 Experimental Design

Our experiment relies on a between-subjects design to elicit subject behavior and beliefs about norms. We conduct two different experiments - a *choice experiment* and a *norms elicitation experiment* - with two different sets of subjects. Subjects in the *choice experiment* do not participate in the *norms elicitation experiment*, and vice-versa.

1.3.1 Choice experiment

We first discuss our *choice experiment*. Following the idea that frames can change behavior by evoking a social identity and its associated norms, we vary whether subjects are shown neutrally-framed or tax-framed dictator games. This treatment is designed to evoke a U.S. political identity (Democratic or Republican) within our subjects.[8]

[8]We target these two political social identities because political identity is a "homegrown" identity (i.e., one that subjects bring with them to the laboratory) that U.S. subjects tend to have internalized by the time they reach adulthood. *Kranton et al.* (2016) review several different approaches to

We deliberately select a frame on which the two political parties strongly differ: tax redistribution. This frame is chosen based on previous empirical work examining the impact of frames on behavior that differs across political party platforms. For instance, the 2012 and 2016 Democratic National Platforms, in multiple separate instances, advocate for the "wealthiest taxpayers to pay their fair share." By contrast, the 2012 and 2016 Republican Platforms "reject the use of taxation to redistribute income." Similarly, a Pew Research Center/USA TODAY survey conducted in January of 2014 shows that, for the question "How much should the government do to reduce the gap between the rich and everyone else," 88% of liberal Democrats answer "A lot" or "Some," compared to only 40% of conservative Republicans.

These platform differences are what we use to construct the tax frame. In the tax-framed treatment, we characterize the dictator game as a wealth redistribution decision, the endowments as initial wealth, and the allocation as a government transfer initiated through the subject's choice. The wording of the tax-framed treatment is:

> *In this economy your wealth is X token(s) and your match's wealth is Y token(s). Use the slider to indicate whether you want the government involved and how large or small the redistribution should be.*

By contrast, the wording of the neutrally-framed treatment is:

> *For this decision you own X token(s) and the other person owns Y token(s). You have the opportunity to give any amount of your X token(s)*

studying homegrown versus lab-created identities. Not only do most U.S. adults possess a political identity, but this identity also exerts high influence on their choices during the decision-making process. *Iyengar and Westwood* (2015) find that the impact of political identity on judgment and behavior exceeds even that of racial identity. *Pickup et al.* have several papers examining how political identities affect voting through social norms. They find that voters are willing to pay a personal cost (vote against their own interests) in order to comply with the norms of their political identity (*Pickup et al.*, 2016, 2018b), and that the personal cost causes the social norms to be strengthened (*Pickup et al.*, 2018a). In our study, we restrict our subjects to U.S. citizens and allow subjects to participate in only one of the treatments.

to the other person or to take any amount of the Y token(s) from the other person for yourself.[9]

Within each treatment, subjects make eleven dictator game decisions. For each dictator game, there are a total of 10 tokens to split between the dictator and a receiver. The eleven dictator games reflect the eleven possible ways to split the initial 10-token endowment, from a situation where the dictator starts with 10 tokens and the receiver starts with none (the standard dictator game), to a case where the dictator starts with no tokens and the receiver starts with all 10. Thus, our initial endowments vary within each subject. We vary the initial endowment because, based on the party platforms, we expect that the endowments will impact dictator choices differently for subjects who identify as Democrats or Republicans: In order to achieve equal allocations, a Democrat will be *willing* to give or take wealth depending on the initial endowment, while a Republican will be *unwilling* to give or take wealth regardless of the initial endowment. Because we know the party platforms, we can make predictions about how changes to initial endowments will affect behavior for Republicans and Democrats.

There is a stream of experimental work examining the effect of varying the initial endowment in the dictator game. Most of this work uses a between-subjects treatment that places the entire initial endowment with the dictator (in one treatment) or with the recipient (in the other treatment) and finds that dictator behavior is not affected by who starts with the endowment (*Dreber et al.*, 2013; *Grossman and Eckel*, 2015; *Halvorsen*, 2015; *Hauge et al.*, 2016; *Goerg et al.*, 2017). On the other hand, *Grossman and Eckel* (2012) and *Krupka and Weber* (2013a) use a between-subjects design comparing a standard dictator game, where the entire endowment starts with the dictator, to a non-standard game, where the initial endowment is divided equally

[9]For the situation where the subject is endowed with all 10 tokens, the subject reads: "You have the opportunity to give any amount of your 10 tokens to the other person." For the situation where her receiver is endowed with all 10 tokens, she instead reads: "You have the opportunity to take any of the 10 tokens from the other person."

between the dictator and the recipient. These papers find that changing the initial endowment distribution causes a significant change in behavior, suggesting that non-extreme initial endowment comparisons might result in behavioral changes. There is also some evidence that sequentially exposing subjects to different initial endowments affects dictator behavior (*Visser and Roelofs*, 2011; *Korenok et al.*, 2014). These systematic differences could explain why variation in the initial endowments sometimes does and sometimes does not affect dictator behavior.[10] However, these papers are not easily amenable to comparison. In the discussion section, we present evidence that shows that our dictators' significant behavioral differences are largely due to our within-subjects design rather than our inclusion of non-extreme initial endowment distributions.

We first administer the games for each group. Then, after subjects complete the decision making rounds, we administer a 5-item demographic questionnaire which is the same regardless of treatment. The questionnaire elicits the degree to which each subject self-identifies as a Republican or a Democrat by asking the question "In politics, as of today, do you consider yourself:" with a response scale that includes the choices "A Republican," "Leaning more towards the Republican Party," "Leaning more towards the Democratic Party," and "A Democrat."[11] In our analysis, a subject's response to this question determines the subject's political identity. Thus, when we refer to a "tax-framed Republican," we are referring to a subject who both is in our tax-framed treatment and self-identifies as a Republican/leaning Republican.

[10]There are also studies showing that gender (*Kettner and Ceccato*, 2014; *Chowdhury et al.*, 2017), stake size (*Leibbrandt et al.*, 2015), or social norm interventions (*Farrow et al.*, 2018) can interact with the initial endowment to change dictator behavior. A separate stream of experimental work has shown that increasing the dictator's choice set to include taking options changes behavior (*List*, 2007; *Bardsley*, 2008; *Bosman and Van Winden*, 2002; *Eichenberger and Oberholzer-Gee*, 1998; *Swope et al.*, 2008; *Zhang and Ortmann*, 2012; *Cappelen et al.*, 2013). For related work in a VCM setting, see *Andreoni* (1995); *Dufwenberg et al.* (2011b); *Grossman and Eckel* (2012); *Brewer and Kramer* (1986); *McCusker and Carnevale* (1995); *Sell and Son* (1997); *Sonnemans et al.* (1998); *van Dijk and Wilke* (2000); *Brandts and Schwieren* (2009).

[11]This question is adapted from Gallup's standard party identification question, in use since 1944 (*Gallup*, 1991).

Upon completion of the questionnaire, each subject is randomly paired with another subject. A random dictator game is then selected for each pair, and a random subject in each pair is selected to be the dictator. That dictator's decision is then implemented.

1.3.2 Norms elicitation experiment

In addition to our *choice experiment*, we conduct a *norms elicitation experiment* with a different set of subjects. In our *norms elicitation experiment*, these subjects are randomly assigned to treatments in which dictator games are described using either a neutral or tax frame. This experiment differs from the *choice experiment* in that it elicits subjects' beliefs about social norms rather than asking them to make redistribution choices.

To elicit social norms, we follow the procedures developed in *Krupka and Weber* (2013a). That is, we describe a specific dictator game and a specific action and ask subjects to rate the "social appropriateness" of that action in that game. For example, we describe a scenario where a dictator is endowed with 10 tokens (the standard dictator game) and transfers 0 tokens to the recipient. In this case, the subject is asked to judge the appropriateness of this action using the following rating scale: "very socially appropriate," "socially appropriate," "somewhat socially appropriate," "somewhat socially inappropriate," "socially inappropriate," and "very socially inappropriate." This six-category scale follows that of *Krupka et al.* (2016). The subject is asked to make these judgments as part of a coordination game in which she is paid if her rating of the appropriateness of the action matches that of another random subject.

Krupka and Weber (2013a) provide evidence that collectively-recognized social norms create focal points in a matching game (see also *Goerg and Walkowitz* 2010; *Schelling* 1980; *Mehta et al.* 1994; *Sugden* 1995). Here, subjects have an incentive

to anticipate and match how others will rate an action as socially appropriate or inappropriate.[12] If there is a social norm that some actions are more or less socially appropriate, respondents are expected to draw on this shared perception in their attempts to match others' ratings.

In our *norms elicitation experiment*, we limit the presentation of scenarios to three: where the dictator is initially endowed with 10 tokens, with 5 tokens, and with 0 tokens. For each scenario, subjects play the ratings coordination game for each of the eleven possible actions (dictator allocates from 0 to 10 tokens for herself). Thus, they play a total of 11 coordination games in each scenario before moving to the description of the next scenario.[13]

[12]Others have adapted the procedures in *Krupka and Weber* (2013a) to elicit norms for a variety of games. For example, *Kimbrough and Vostroknutov* (2016), *Gächter et al.* (2017), *Veselý* (2015), *Erkut et al.* (2015), *D'Adda et al.* (2015), *Gangadharan et al.* (2016), and *Banerjee* (2016) examine norm compliance across a variety of games using the Krupka and Weber norm elicitation protocol. However none of these studies examines identity-dependent norms. Yet a different approach to eliciting norms is to use third party advisors (*Schram and Charness*, 2011); however, this approach is more challenging to adapt to the study of identity-dependent norms. Another similar alternative is used in *Bicchieri and Chavez* (2010), where norms are elicited by asking proposers and responders in an Ultimatum Game to guess how many responders perceive each of the proposers' options as a fair option. However, this approach would limit what we would be able to say about a "set" of appropriate actions.

[13]Subjects read about each of these three scenarios, but the order in which they read about them is randomized. In total, each subject in the *norms elicitation experiment* plays 33 coordination games.

1.3.3 Experimental Procedure

Our subjects are workers from Amazon Mechanical Turk (MTurk).[14] Workers on MTurk perform small tasks set by requesters, who then pay the workers for completing the tasks. For economics experiments, workers are paid a standard flat rate plus a bonus which depends on their actions in the experiment. Requesters also pay Amazon a 20% commission for completed tasks. In this sense, the flat rate corresponds to a show-up fee, the bonus corresponds to incentives, and the commission corresponds to fees one might pay to use a lab in a traditional economics laboratory experiment.

In our *choice experiment*, subjects first complete an unincentivized 10-item questionnaire. [15] They then proceed to the dictator games. Figure 1.1 presents a screenshot of a dictator decision that tax-framed subjects encounter. The depicted decision is one where the initial endowment for the dictator is 8 tokens and for the receiver is 2 tokens.[16] The dictator indicates her decision by moving the white box along the slider (in Figure 1.1, the slider has already been moved to indicate a transfer of 8 tokens to the receiver). The subject cannot move on to the next screen until she

[14]MTurk was started in 2005 as a spot market for labor. It is now commonly used for experimental research. The population of MTurk workers is at least as representative of the U.S. population as traditional subject pools and several classic experiments have been replicated online such as the prisoner's dilemma, priming, and framing experiments (*Horton et al.*, 2011; *Chandler and Kapelner*, 2013; *Paolacci et al.*, 2010). Further, *Huff and Tingley* (2015) compare individual and political characteristics of MTurkers against respondents of the Cooperative Congressional Election Survey and find that the groups are largely similar. Although MTurk workers take on many tasks (often working for two hours a day on such tasks), it is unlikely that they will have encountered the norms rating activity in previous tasks because the norms rating activity has not yet been used in an online setting. It is possible that they have encountered the dictator game before and may have "set" or "routine" responses to such games. However this is less concerning because our treatments vary the tax frame rather than the task. So, if we observe that tax-framed subjects behave differently from neutrally-framed subjects on the same task, we can still attribute this change in behavior as being due to the effect of the frame.

[15]The unincentivized questionnaire has three components. First, subjects are shown five pairs of headshots and asked to choose the more attractive one in each pair. Second, subjects are asked to look at two pictures of people in lines and to choose which line is longer. Finally, subjects are asked to guess the average temperature in five states in a previous year. These questions were part of a larger study which is unrelated to the research questions of this paper.

[16]The order in which subjects encounter the eleven situations is randomized according to four blocks. The four blocks have the following order: in block (1) the dictator's initial endowment varies from 0, 1, 2, ..., 10 tokens; in block (2) it varies from 5, 0, ..., 4, 6, ..., 10 tokens; in block (3) it varies from 10, 9, 8, ..., 0 tokens; and in block (4) it varies from 5, 10, ..., 6, 4, ..., 0 tokens.

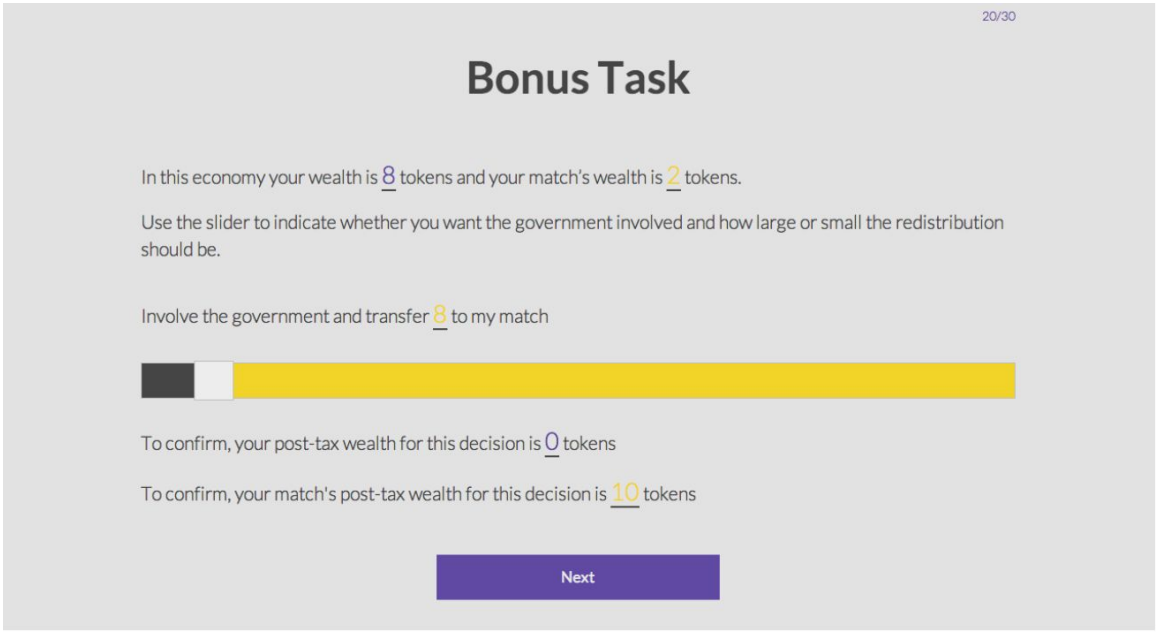

Figure 1.1: Screenshot of the *choice experiment* redistribution task for tax-framed subjects. The white slider element starts in the "neutral" position which is located either to the left or to the right of the slider (this is randomized). The slider must be moved off of the neutral position for the subject to indicate her choice (the slider depicted here has already been moved). The numbers on the screen also dynamically update as the slider is moved to reflect the action being taken and the outcome of that action.

actively moves the slider. The neutral position of the slider is left/right randomized for each decision. Once the dictator begins to move the white box along the slider, the other elements of the screen dynamically update to reflect the choice being made as well as the final allocation.

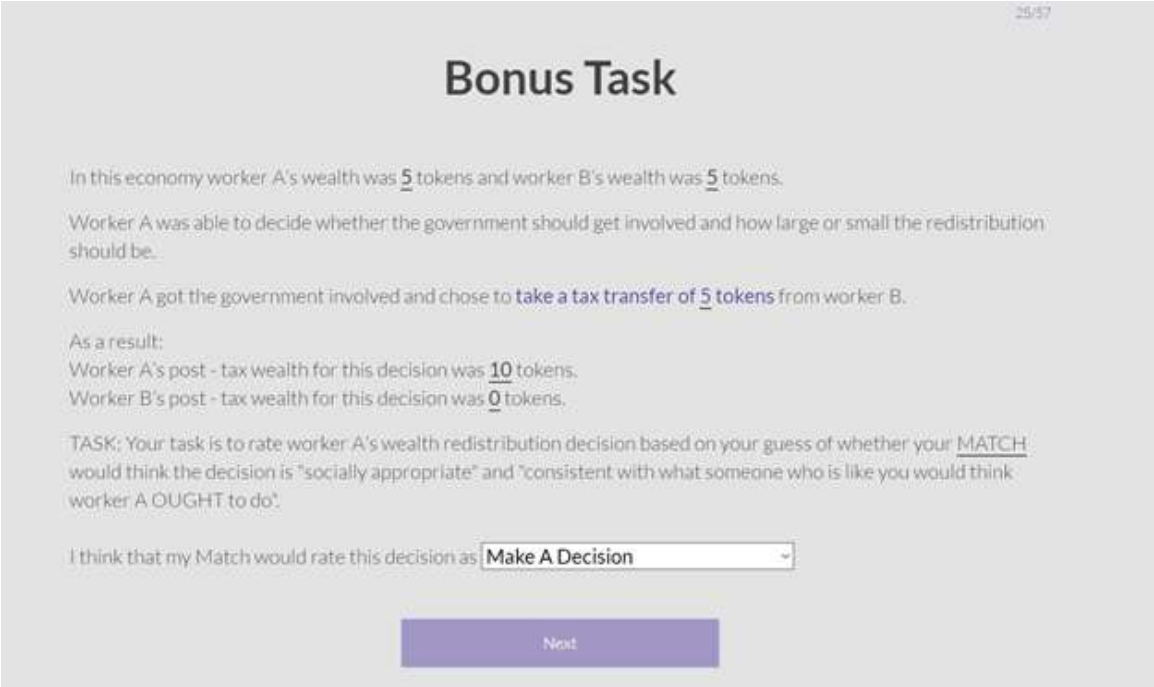

Figure 1.2: Screenshot of the *norms elicitation experiment* ratings task for tax-framed subjects.

Figure 1.2 presents a screenshot from the *norms elicitation experiment* of a situation where the dictator's initial endowment is 5 tokens and the dictator's chosen action is "take 5 tokens." For example, a subject in the tax-framed treatment reads about this situation and guesses how appropriate another MTurker would rate the action "take a tax transfer of 5 tokens from worker B." Using the drop-down menu, the subject indicates her guess of how "socially appropriate" and "consistent with

what someone who is like you would think worker A OUGHT to do."

Table 1.1 presents the number of Democratic and Republican subjects in each treatment. On average, subjects in the *choice experiment* and *norms elicitation experiment* receive $1.00 and $1.34, respectively, for their participation. We conduct the experiment between 2014 and 2016.[17]

Table 1.1: Number of Democratic and Republican subjects in each treatment

	Neutrally-framed	Tax-framed
Norms elicitation experiment	Republicans: 65 Democrats: 114	Republicans: 68 Democrats: 132
Choice experiment	Republicans: 73 Democrats: 154	Republicans: 130 Democrats: 270

1.4 Results

We begin our discussion of the results by examining norms and behavior using the data from our *norms elicitation* and *choice experiments*. We then present evidence that the social identity model elucidates the mechanism, social norms, through which frames affect choice.

1.4.1 Norms and behavior

1.4.1.1 Analysis of norms

To study the effect of frames on norms, we follow *Krupka and Weber* (2013a) and transform the appropriateness ratings from the *norms elicitation experiment* into an empirical measure of the norm by converting subjects' ratings into numerical scores (or norm ratings). Specifically, a rating of "very socially inappropriate" receives a score of -1, "socially inappropriate" receives a score of -0.6, "somewhat socially inappropriate" receives a score of -0.2, "somewhat socially appropriate" receives a score

[17]The full experimental instructions are available in the Appendix.

of 0.2, "socially appropriate" receives a score of 0.6, and "very socially appropriate" receives a score of 1.[18]

To empirically estimate Democratic (Republican) tax-framed norms when the dictator's initial endowment is 10 tokens, we restrict our analysis to responses from subjects in the tax-framed treatment who (1) self-report that they are Democrats (Republicans) and (2) rate the situation where a dictator has an initial endowment of 10 tokens. As in *Krupka and Weber* (2013a), we take the average norm rating for each action. We repeat this process for initial endowments of 5 and 0 tokens to obtain empirical proxies for the Democratic (Republican) tax-framed norms for the respective endowments. Similarly, we construct neutrally-framed norm profiles for Democrats (Republicans) using the responses from subjects in the neutrally-framed treatment who self-report that they are Democrats (Republicans).

First, we restrict our attention to the neutral frame. Some previous literature suggests that changing the initial endowment distribution (such that the dictator no longer holds all of the initial endowment but retains the rights to determine the final allocation) causes the dictator to change her behavior. *Krupka and Weber* (2013a) show that this stems from changes in the norms. Thus, we predict that there will be differences in normative ratings across endowment distributions.

Hypothesis 1 (**Norms: endowments affect norms in the neutral frame**)**.** *Norm ratings will differ across initial endowment distributions in the neutral frame.*

Figure 1.3 displays the average norm ratings for the three initial endowments (0, 5, or 10 tokens) for each identity and frame combination. The x-axis reflects the number of tokens the dictator allocates to herself (e.g., the dictator choice to allocate 0 to

[18]Note that this transformation is also used in *Kimbrough and Vostroknutov* (2016), *Gächter et al.* (2013), *Veselỳ* (2015), *Erkut et al.* (2015), *D'Adda et al.* (2015), *Gangadharan et al.* (2016), *Banerjee* (2016), and *Gächter et al.* (2017). By giving the ratings a numerical value, we are imposing ratio scale characteristics on measurements that are, by design, ordinal. In some of what follows this is merely for convenience, such as when we use a rank-order test for the equality of distributions. However, in other situations, it implicitly adds extra assumptions upon which our analysis is then conditional, such as when we compare means.

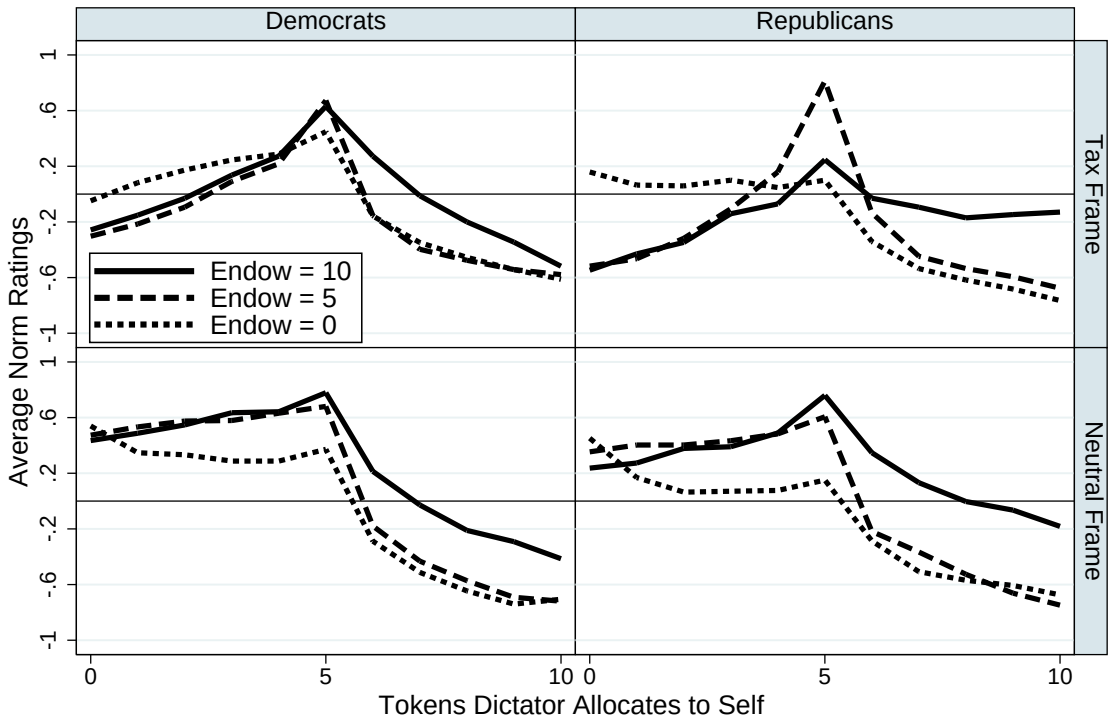

Figure 1.3: Average norm ratings by frame, initial endowment, identity, and dictator action.

herself and 10 to her match is depicted as "0" on the x-axis). Note that the choice that a dictator must make to achieve the same final allocation (e.g., "0 to self") differs by the initial endowment. The y-axis reflects the values that the average norm ratings may take, with -1 representing the rating for "very socially inappropriate" and 1 representing the rating for "very socially appropriate."[19]

The norm ratings depicted in Figure 1.3 support Hypothesis 1 by showing that neutrally-framed norms differ across initial endowments for both Democrats and Republicans. In order to formally test Hypothesis 1, we run 33 Wilcoxon signed-rank tests (with Bonferroni correction). Each subject in the *norm elicitation experiment* gives a norm rating for each of the initial dictator endowments of 0, 5 and 10 tokens, as well as each of 11 possible dictator actions, for a total of 33 norm ratings. For each Wilcoxon signed-rank test, we treat each subject-action pair as one observation with two linked norm ratings and run a separate test for each pair of endowments and each action. These tests lead to the following result:

Result 1 (**Norms: endowments affect norms in the neutral frame**). *Norm ratings differ significantly across initial endowment distributions in the neutral frame.*

Support. The p-values of the 33 Wilcoxon signed-rank tests are displayed in Table A.3. When applying a Bonferroni correction for multiple hypothesis testing, the p-value threshold for significance at the 5% level is 0.0015. Comparing the norm ratings for initial dictator endowments of 0 and 5 tokens, 6 out of 11 comparisons are significant at the 5% level. Comparing the norm ratings for initial dictator endowments of 0 and 10 tokens, 9 out of 11 comparisons are significant at the 1% level. Compar-

[19]We note that our results share several important features with *Krupka and Weber* (2013a) as well as with *Kimbrough and Vostroknutov* (2016). Both papers used student data while we use MTurk data. We all find a peak in norm ratings at the (5,5) or equal split action; we all have a steep drop off as the dictator keeps more for herself; we all have the most negative rating for the most selfish action (dictator keeps all for herself); we all have a less steep drop off for actions where the dictator gives more to the recipient than she keeps for herself. However, we also note that our data show a much less steep decline on either side of the peak than the previous papers. This may be due to differences in subject pool, stake size of the scale on which norms ratings are elicited.

ing the norm ratings for initial dictator endowments of 5 and 10 tokens, 6 out of 11 comparisons are significant at the 5% level.

We find that our subjects' norms are significantly affected by the initial endowment distribution. Notably, Figure 1.3 shows that people consider keeping all of the tokens to be more appropriate when the dictator starts with all of the tokens than when the receiver starts with some tokens. Also, taking any of those tokens is seen as less appropriate when the receiver starts with all of the tokens than when the dictator starts with all of the tokens. The status quo has a normative advantage over other outcomes.

Continuing to restrict our attention to the neutral frame, we next compare the norm ratings of Democrats and Republicans. When the redistribution task is neutrally framed, we expect no variation in norm ratings because the identities (and associated norms) are not made salient by the frame. This leads to the following hypothesis:

Hypothesis 2 (**Norms: Democratic and Republican norms in the neutral frame**)**.** *Norm ratings will not differ between Democrats and Republicans in the neutral frame.*

Figure 1.3 also supports Hypothesis 2 by showing that Democratic and Republican norm ratings are very similar in the neutral frame. To test Hypothesis 2, we run several Mann-Whitney U tests between the Democratic and Republican norm ratings, one for each endowment-action pair. This gives us 33 Mann-Whitney U tests, and leads to the following result:

Result 2 (**Norms: Democratic and Republican norms in the neutral frame**)**.** *Norm ratings do not differ significantly between Democrats and Republicans in the neutral frame.*

Support. The p-values of the 33 Mann-Whitney U tests are displayed in Table A.4. When applying a Bonferroni correction for multiple hypothesis testing, the p-value threshold for significance at the 10% level is 0.0030. Since the lowest p-value in Table A.4 is 0.0039, none of the norm ratings are significantly different at the 10% level.

When the redistribution tasks are neutrally framed, Democrats and Republicans agree on what constitutes appropriate dictator behavior. This remains true across different initial endowment distributions. Without rhetoric, we agree on what is and is not appropriate, because our political identities are not made salient.

Finally, we compare norm ratings across frames. Frames change the situation by evoking a person's social identity and the associated identity-dependent norms. For this reason we anticipate that the social norms for the tax-framed treatment will differ from those for the neutrally-framed treatment, for both identities. This leads to the following hypothesis:

Hypothesis 3 (**Norms: frames affect norm ratings**). *For each identity, the tax-framed norm ratings will differ from the neutrally-framed norm ratings.*

The norm ratings depicted in Figure 1.3 also support Hypothesis 3 by showing that tax-framed norms differ from neutrally-framed norms for each identity. For Democrats, the tax frame seems to make an equal split more appropriate compared to other actions, particularly with respect to giving more than half of the tokens to the receiver. For Republicans, the tax frame seems to make the status quo more appropriate, seen most clearly when the dictator starts with all of the tokens. In that scenario, the dictator keeping all of the tokens becomes more appropriate than the dictator giving away all of the tokens. For both Republicans and Democrats, the impact of the tax frame is in line with the respective party platforms.

To test Hypothesis 3, we again run several Mann-Whitney U tests between the tax- and neutrally-framed norm ratings, one for each endowment-action pair, separated

by identity. This gives us 66 Mann-Whitney U tests, 33 for each identity, and leads to the following result:

Result 3 (**Norms: frames affect norm ratings**). *Except when the dictator's initial endowment is 0 tokens, norm ratings differ between the tax and neutral frames.*

Support. The p-values of the 66 Mann-Whitney U tests are displayed in Table A.5. When applying a Bonferroni correction for multiple hypothesis testing, the p-value thresholds are 0.00152, 0.00076, and 0.00015 for significance at the 10%, 5%, and 1% levels, respectively. When the dictator's initial endowment is 0 tokens, only 1 out of 11 comparisons for the Democratic norms (when they keep 0 tokens) shows significant differences at the 10% level between frames. For Republican norms, 0 out of 11 comparisons are significant at the 10% level between frames. For the other two initial endowments, at least 5 of 11 actions show significant differences at the 1% level for each identity, mainly for the dictator keeping fewer tokens. The tax frame significantly reduces the appropriateness for these actions.

1.4.1.2 Analysis of dictator choice

We next examine the results from our *choice experiment*. First, we present three hypotheses and results that mirror those for norms. These are ex ante hypotheses, predicting behavior before we have the norms results.

Our first choice hypothesis is motivated by the previous literature. Some studies that use a between-subjects design with regards to initial endowment distribution (e.g. *Grossman and Eckel*, 2012; *Krupka and Weber*, 2013a) have found that dictator behavior differs significantly depending on whether the dictator or recipient holds the initial endowment.[20] Thus, focusing on the neutral frame, we have the following

[20]We acknowledge that much of the previous literature on this topic finds no effect of the dictator's initial endowment on the amount they keep. We address this further in the Discussion section.

hypothesis:

Hypothesis 4 (Choice: endowments affect choice in the neutral frame). *Dictator choice will differ across initial endowment distributions in the neutral frame.*

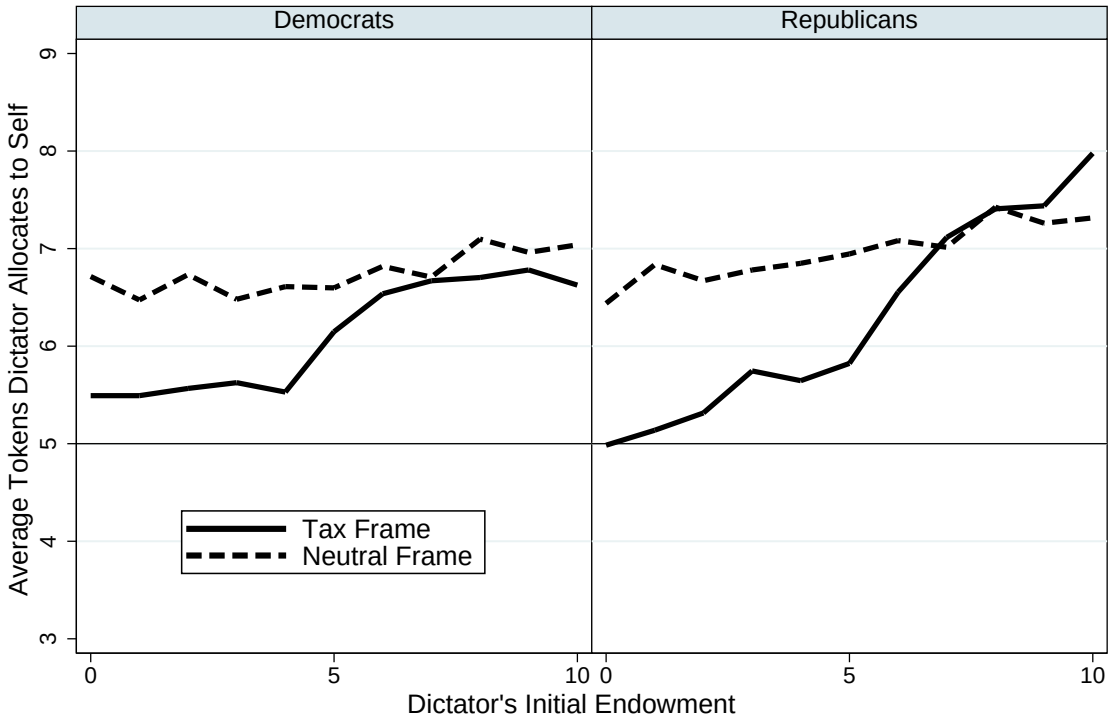

Figure 1.4: Average dictator choices by frame, initial endowment, and identity.

Figure 1.4 displays the average dictator decision in the *choice experiment* for each endowment, separated by frame treatments and political identity. In Figure 1.4, the average choices of tax-framed dictators are indicated by the solid lines while the average choices of neutrally-framed dictators are indicated by the dashed lines. For both Democrats and Republicans, the dashed lines are upward sloping, supporting Hypothesis 4.

To test Hypothesis 4, we run an OLS regression of the number of tokens the dictator keeps in the neutral frame on the endowment, clustered at the individual level. This leads to the following result:

Result 4 (**Choice: endowments affect choice in the neutral frame**). *Dictators keep significantly more tokens when they are initially endowed with more tokens.*

Support. Column 1 of Table 1.2 shows the results of the OLS regression of number of tokens kept by the dictator on the dictator's initial endowment. The coefficient on the initial endowment is 0.059, significant at the 1% level, indicating that dictators keep about 0.6 more tokens when they start with 10 tokens, compared to when they start with no tokens.

Table 1.2: OLS regressions of tokens kept by dictators in the neutral frame

Dependent variable: tokens kept by dictator		
	(1)	(2)
Republican		0.05
		(0.390)
Endowment	0.06***	0.05***
	(0.016)	(0.018)
Republican × Endowment		0.03
		(0.035)
Constant	6.52***	6.51***
	(0.181)	(0.218)
Observations	2,497	2,497
R^2	0.004	0.006

Notes: 1. Standard errors (in parentheses) are adjusted for clustering at the individual level.
2. Significant at the *** 1 percent, ** 5 percent, and * 10 percent levels.

Next, we examine how Democratic and Republican dictators behave differently from each other in the neutral frame. In the neutral frame, political identities are not salient, so we expect no variation in behavior that stems from those identities. Thus, we predict that there will be no differences between Democratic and Republican behavior in the neutral frame, giving us the following hypothesis:

Hypothesis 5 (**Choice: Democratic and Republican dictator behavior in the neutral frame**). *Democratic and Republican dictators in the neutral frame will*

keep the same number of tokens for the same initial endowment.

We can see in Figure 1.4 that in the neutral frame (dashed lines), the Democrats and Republicans seem to keep a similar number of tokens for the same initial endowment. To test Hypothesis 5, we run an OLS regression of the number of tokens kept by the dictator on the initial endowment, whether the subject is a Republican, and an interaction term, clustered at the individual level. This gives us the following result:

Result 5 (**Choice: Democratic and Republican dictator behavior in the neutral frame**). *Democratic and Republican dictators in the neutral frame show no significant differences in the number of tokens that they keep for a given endowment.*

Support. Column 2 of Table 1.2 shows the results of the OLS regression of number of tokens kept on the initial endowment and the dictator's political identity. The coefficients on both "Republican" and the interaction term of "Republican" with "Endowment" are both insignificant at the 10% level.

As with our norm results, political identity does not affect dictator behavior when those political identities are not made salient through framing. Without rhetoric, people make similar choices regardless of political identity.

Next, we examine how the frame affects dictator choice. Since we expected that framing would affect the norms, we also expect that framing will affect dictator choice. This leads to the following hypothesis:

Hypothesis 6 (**Choice: frames affect choice**). *For a particular identity, the allocation choices of tax-framed dictators will differ from those of the neutrally-framed dictators.*

From Figure 1.4, we can see that both Democratic and Republican dictators in the tax-framed treatment are more responsive (have steeper slopes) to their initial endowment, relative to dictators in the neutrally-framed treatment.

In Table 1.3, we present the results from OLS regressions to determine whether differences across frames are significant for either Democrats (column 1) or Republicans (column 2). In particular, we regress the number of tokens kept by the dictator on a dummy for the frame ("Tax-framed" is 1 for subjects in the tax-framed treatment and 0 for subjects in the neutrally-framed treatment), the dictator's initial endowment ("Endowment"), and their interaction term ("Tax-framed × Endowment").

Table 1.3: OLS regressions testing effect of frame and endowment on dictator choice (by identity)

Dependent variable: tokens kept by dictator		
	(1) Democrats	(2) Republicans
Tax framed	-1.19***	-1.82***
	(0.275)	(0.457)
Endowment	0.05**	0.08**
	(0.018)	(0.030)
Tax framed × Endowment	0.11***	0.23***
	(0.027)	(0.049)
Constant	6.51***	6.55***
	(0.218)	(0.324)
Observations	4,664	2,233
R^2	0.035	0.084

Notes:
1. Standard errors (in parentheses) are adjusted for clustering at the individual level.
2. Significant at the *** 1 percent, ** 5 percent, and * 10 percent levels.
3. The number of observations in column 1 comes from the decisions of 424 Democrats in each of 11 dictator games. The number of observations in column 2 comes from the decisions of 203 Republicans in each of 11 dictator games.

Taking the visual and regression evidence together, we obtain the following result:

Result 6 (Choice: frames affect choice). *Subjects playing the tax-framed dictator game allocate their initial endowments differently than do subjects playing the neutrally-framed dictator game.*

Support. The results in Table 1.3 show that the coefficients for "Tax-framed" and "Tax-framed × Endowment" are significant ($p < 0.01$) for both Democrats (column 1) and Republicans (column 2).

When the tax frame is introduced, dictators act significantly differently from when there is a neutral frame. This is again related to the finding that norms are affected by the framing.

1.4.2 Testing the social identity model

To test whether the *social identity model* can explain our results, we first show that the data from our *norms elicitation experiment* improves our ability to account for behavior in our *choice experiment*. Note that this exercise also calibrates the model. We then compare out-of-sample predictions with actual subject behavior in our experiment.

1.4.2.1 Predicting choice using norms

When we take the norms data and the *social identity model* as given, we can predict how individuals will behave based on those norms. We then examine whether our subjects follow those predictions.

The norms displayed in Figure 1.3 show the trade-off between payoffs and norms for different situations. We can think of these norm ratings graphs as budget constraints with dictator preferences represented as the linear indifference curves in equation 1.1. Individual level differences would be represented by different slopes for the indifference curves, with steep indifference curves for dictators who do not care about following social norms, and shallow indifference curves for dictators who do care about social norms. With this formulation, and the specific social norms displayed in Figure 1.3, we can make several predictions about dictator behavior.

a) In the neutral frame, for initial endowments of 5 and 10 tokens, dictators will only keep either 5 or 10 tokens.

b) In the neutral frame, for initial endowment of 0 tokens, dictators will only keep 0, 5, or 10 tokens.

c) In the tax frame, for initial endowment of 5 tokens, dictators will only keep 5 or 10 tokens.

d) In the tax frame, for initial endowment of 10 tokens, Democratic dictators may keep any number of tokens between 5 and 10, while Republican dictators will only keep either 5 or 10 tokens.

e) In the tax frame, for initial endowment of 0 tokens, Democratic dictators will only keep either 5 or 10 tokens, while Republican dictators may keep any number of tokens between 0 and 5, or 10 tokens.

These restrictions on dictator behavior stem from the shapes of the norm ratings in different situations and for different identities. In the neutral frame with initial endowment of 5 or 10 tokens, the most appropriate action is to keep 5 tokens. This fact immediately eliminates keeping any fewer than 5 tokens as utility maximizing, because a dictator who keeps fewer than 5 tokens can gain both monetary- and norm-compliance-utility by keeping 5 tokens instead. The norm graphs are convex (i.e. concave up) between 5 and 10, and this eliminates keeping between 6 and 9 tokens because they are not utility maximizing. This leaves us with two possible utility maximizing choices: dictators keeping 5 or 10 tokens.[21]

In the neutral frame with initial endowment of 0 tokens, the most appropriate action is to keep 0 tokens. Due to the shape of the norm graphs, keeping 5 tokens

[21] Which of these actions is chosen depends on the steepness of their indifference curves. Because we do not have individual level estimates of this steepness, we cannot predict how many dictators will keep 5 tokens or how many will keep 10 tokens.

is also a possible utility-maximizing choice. The norm graphs are convex between keeping 0 and 5 tokens, and between keeping 5 and 10 tokens. Therefore, depending on the steepness of their indifference curves, utility maximization is achieved when the dictator keeps 0, 5 or 10 tokens.

In the tax frame with initial endowment of 5 tokens, the most appropriate action is to keep 5 tokens for both Democrats and Republicans, so the prediction here is the same as in the neutral frame, that both Democratic and Republican dictators will only keep either 5 or 10 tokens.

In the tax frame with initial endowment of 10 tokens, the Democratic norm graph between keeping 5 and 10 tokens is linear, so it is possible that some dictators will keep between 6 and 9 tokens. For the Republican norm graph, keeping 5 tokens is the most appropriate action and it is convex between keeping 5 and 10 tokens, so the dictators will only keep either 5 or 10 tokens.

Finally, in the tax frame with initial endowment of 0 tokens, the Democratic norm graph is similar to the case when the endowment is 5 tokens, so the dictators will choose to keep either 5 or 10 tokens. For Republicans, the norm graph between keeping 0 and 5 tokens is linear, so it is possible that some dictators will keep between 1 and 4 tokens.

Figure 1.5 displays the histograms of dictator actions in each situation and for each identity. Because there is no obvious null hypothesis, we use these distributions of dictator choice to characterize whether behavior generally follows the above identified patterns.

a) In the neutral frame, for initial endowments of 5 or 10 tokens, 80% of dictators keep 5 or 10 tokens.

b) In the neutral frame, for initial endowment of 0 tokens, 82% of dictators keep 5 or 10 tokens, and 5% keep 0 tokens.

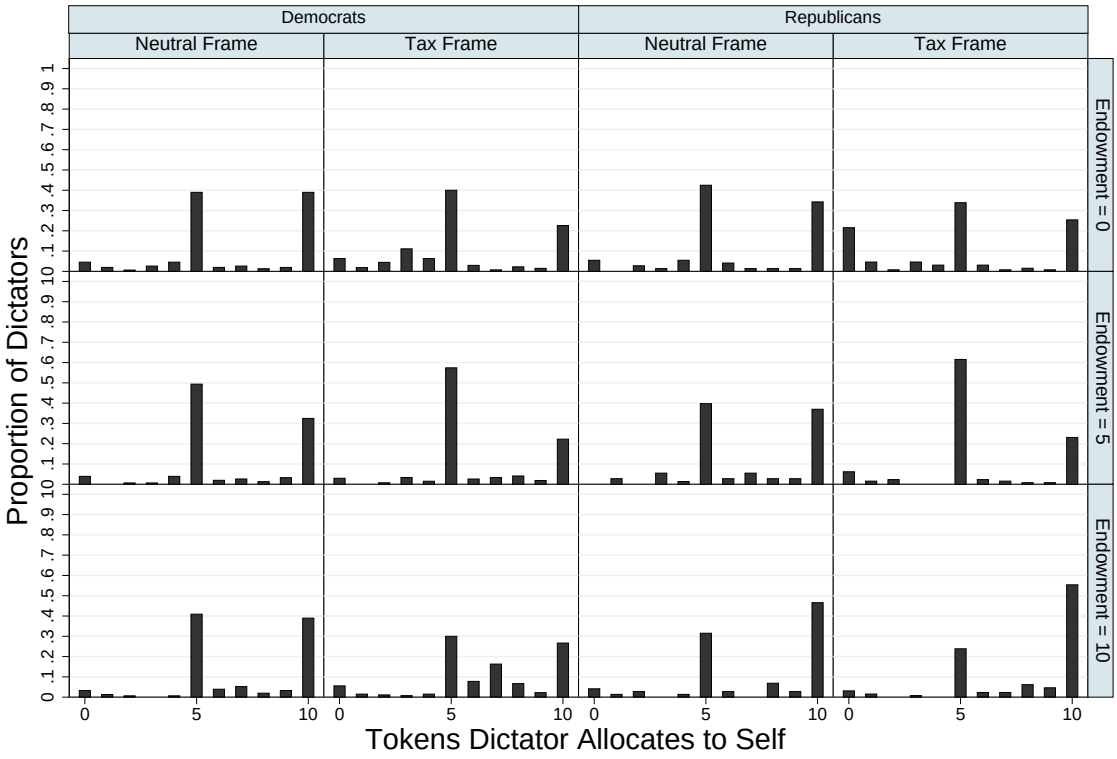

Figure 1.5: Dictator choices by frame, initial endowment, and identity.

c) In the tax frame, for initial endowment of 5 tokens, 81% of dictators keep 5 or 10 tokens.

d) In the tax frame, for initial endowment of 10 tokens, 57% of Democratic dictators keep 5 or 10 tokens, and 33% keep between 6 and 9 tokens. 79% of Republican dictators keep 5 or 10 tokens.

e) In the tax frame, for initial endowment of 0 tokens, 63% of Democratic dictators keep 5 or 10 tokens. 59% of Republican dictators keep 5 or 10 tokens, 22% keep 0 tokens, and 13% keep between 1 and 4 tokens.

For most of the cases where the norm ratings predict that dictators will only keep 5 or 10 tokens, at least 79% of dictators follow that prediction. Also, in cases where the norm ratings predict behavior other than keeping 5 or 10 tokens, we see that there are a substantial number of dictators who follow those predictions. Perhaps most striking are the tax-framed Republicans who have an initial endowment of 0 tokens; 22% of these dictators keep 0 tokens. This behavior is difficult to justify without appealing to norms.

There are some cases where the norm ratings predict behavior we do not observe. While we predict that some dictators in the neutral frame with an initial endowment of 0 tokens would keep 0 tokens, only 5% of them do so. Also, for tax-framed Democrats with an initial endowment of 0 tokens, 30% of dictators keep between 0 and 4 tokens, behavior not predicted by the shape of the norm ratings.

The *social identity model* seems to be consistent with the norms and dictator behavior that we observe. In the following section, we compare the *social identity model* to the *standard model* (which only takes into account an individual's own payoff) as a benchmarking exercise. However, other models might also be able to account for the observed dictator behavior. For example, an inequality aversion model such as the one proposed by *Fehr and Schmidt* (1999) or a social preferences model

such as the one proposed by *Charness and Rabin* (2002) could predict that dictators mostly give either half or none of the total endowment to the receiver. For *Charness and Rabin*'s *social preferences model*, this would depend on each individual's other-regarding preferences. We can estimate the key parameters of the *social preferences model* using our data, and it encompasses *Fehr and Schmidt*'s inequality aversion model. As such, in the next section we compare the relative ability of the *social identity model* to account for behavior by comparing it to both the *standard* and *social preferences models*.

1.4.2.2 Comparing choice models

In all models we consider, we assume that an individual i employs a logistic choice rule, where her probability P_i of choosing any action a^j depends on the relative expected utility $u_i(a^j)$ of that action compared to other actions:

$$P_i(a_i = a^j) = \frac{\exp(u_i(a^j))}{\sum_k \exp(u_i(a^k))} \tag{1.2}$$

Our first specification assumes that utility depends only on the dictator's own payoff. This is equivalent to setting $\gamma_i = 0$ in Equation 1.1 (i.e. the person does not care about complying with the social norms for her identity). To estimate the weight placed on monetary payoffs, we then impose a linear restriction on $V_i(\cdot)$, such that, for any final payoff, k, $V_i(k_i) = \beta_i k_i$. Thus, we estimate the weight, β_i, that an individual places on the money she receives from a particular choice as follows:

$$u_i(k_i) = \beta_i k_i \qquad \text{(Standard Model)}$$

Our second specification uses the *Charness and Rabin* (2002) social preferences model (hereafter referred to as the *social preferences model*), an important alternative model that could account for the transfers that we observe. This model has become widely adopted, and is now one of the most important alternative models of dictator

behavior, providing an important alternative benchmark in addition to the *standard model*. In the *social preferences model*, an individual's utility includes both her own and her match's payoffs. An individual i's utility is:

$$u_i(k_i) = (\rho \cdot r + \sigma \cdot s)(10 - k_i) + (1 - \rho \cdot r - \sigma \cdot s)k_i$$

(Social Preferences Model)

where $r = 1$ if $k_i > 5$ and $r = 0$ otherwise, and $s = 1$ if $k_i < 5$ and $s = 0$ otherwise.

These first two specifications are outcome-based models. As such, they do not predict different behavior between our tax and neutral frame treatments. If changes in a subject's behavior are driven by changes in norms across initial endowment levels e, and across frames, then the weight that individuals place on complying with the norm, γ_i, should be significantly different from 0. Thus, in our third specification, we assume that an individual is motivated by both the monetary gain from the action as well as the social appropriateness of that action:

$$u_i(k_i, I_i, e) = \beta_i k_i + \gamma_i N(k_i | I_i, e)$$

(Social Identity Model)

The main conceptual difference between the *social identity* and *social preferences models* is that, in addition to receiving utility from their own payoffs, dictators in the former case receive utility from conforming to the identity's norms while in the latter they receive utility from their matches' payoffs.[22] Our tax and neutral treatments offer an important wedge between these models. In our experiment, the *social preferences model* does not make different predictions for behavior between these treatments while the *social identity model* does.

To test which of these models best accounts for identity-dependent choice, we fit individual utility functions to our choice data using conditional logistic regressions.

[22] *Charness and Rabin* (2002) on p. 823 say about their model that "the parameters ρ and σ allow for a range of different 'distributional preferences,' that rely solely on the outcomes...."

Note that in a conditional logistic regression (*McFadden*, 1974) where the dependent variable is the selected action, the variation reflects variation across the characteristics of the possible actions.[23] In our experiment, these characteristics are the payoffs and norms. When we change the initial endowment amount or change the framing, we hold the monetary payoff constant. Thus, the source of variation is the variation in norms.

We separately fit these models for each treatment and identity. While these models do not predict differences between the treatments, estimating the parameters of the models for each treatment separately allows them more flexibility and makes it easier to compare them to the *social identity model*. Also, any observed differences in the estimated parameters of these models would provide evidence against them. We do, however, allow for the possibility that subjects who are Democrats or Republicans have different underlying preferences (*Kranton and Sanders*, 2017). In this way, our approach goes some way toward making both the *standard* and *social preferences models* identity-based models. For the *social preferences model*, this makes it similar, but not equivalent, to the *Chen and Li* (2009) model.[24]

To compare the likelihood that each model fits the observed data, we use the Akaike and Bayesian information criteria (AIC and BIC). Specifically, smaller AIC and BIC values indicate a better fit of the model to the data.[25] Furthermore, since both the AIC and BIC penalize models for the number of parameters, if norms have no influence on behavior, we expect the *social identity model* to have larger AIC and BIC values than the *standard model*. This leads to the following prediction:

[23] Conditional logistic models are similar to multinomial logistic models. However, conditional logistic models emphasize the characteristics of the alternatives, while multinomial logistic models depend on the characteristics of the individual making the choice. See *Hoffman and Duncan* (1988) for a comparison of these models.

[24] *Chen and Li* (2009) allow for different social preference parameters depending on the in- and out-group status of the interacting parties. We do not estimate the "guilt" (ρ) and "envy" (σ) parameters as a function of the in- and out-group status of the dictator and the recipient, but rather estimate different parameters for each identity group – the Democrats and Republicans.

[25] A more in-depth discussion of these two estimators can be found in *Aho et al.* (2014).

Hypothesis 7 (**Choice and Norms: social identity model accounts for behavior**). *A model including identity-dependent norms as an explanatory variable for the corresponding behavior should improve our ability to account for behavior (as measured through decreases in AIC and BIC) as compared to models excluding those norms as an explanatory variable.*

Table 1.4 reports the results from several conditional logistic regressions for Democrats (panel A) and Republicans (panel B) using the norms and behavior obtained from our tax-framed subjects in columns 1 to 3 and neutrally-framed subjects in columns 4 to 6.[26] This gives us the following result:

Result 7 (**Choice and Norms: the *social identity model* accounts for subject behavior**). *For each combination of frame and identity, the social identity model better accounts for the observed variations in behavior than either the standard model or the social preferences model.*

Support. The results in Table 1.4 show that the AICs and BICs of the *social identity model* are smaller than those of the other two models.

We first examine the results the tax frame (columns 1 to 3). For Democrats, both the AIC and BIC are smaller for the *social identity model* (3183 and 3197, respectively) than for the *standard* (3789 and 3796, respectively) or the *social preferences models* (3592 and 3607, respectively). Republicans show a similar pattern (AIC = 1496 vs. 1809 and 1805 and BIC = 1509 vs. 1816 and 1818). A Vuong test for each identity comparing the goodness-of-fit between these models shows that the *social identity model* fits better than either the *standard* or the *social preferences models* for the tax-framed treatment ($p < 0.01$ for both Democrats and Republicans).

[26]For these conditional logistic regressions, we do not distinguish whether the decision is made when the initial endowment is 0, 5, or 10 tokens. This is captured by the different average norm ratings attached to each action for each endowment in the *social identity model*, and no differences are predicted for the other two models.

Table 1.4: Conditional logistic estimation using average norm ratings

Panel A: Democrats						
	Tax Frame			Neutral Frame		
	(1) Standard Model	(2) Social Preferences Model	(3) Social Identity Model	(4) Standard Model	(5) Social Preferences Model	(6) Social Identity Model
Own payoff (β)	0.112***		0.419***	0.191***		1.043***
	(0.013)		[0.039]	(0.024)		[0.089]
Other's payoff when ahead (ρ)		0.575***			0.502***	
		(0.022)			(0.031)	
Other's payoff when behind (σ)		0.234***			0.194***	
		(0.029)			(0.054)	
Norms (γ)			2.833***			4.602***
			[0.180]			[0.361]
$0.1 \cdot \frac{\gamma}{\beta}$			0.676***			0.441***
			[0.033]			[0.010]
Observations	8910	8910	8910	5082	5082	5082
Log likelihood	-1894	-1794	-1590	-1032	-995.0	-795.5
AIC	3789	3592	3183	2066	1994	1595
BIC	3796	3607	3197	2072	2007	1608
Panel B: Republicans						
	Tax Frame			Neutral Frame		
	(1) Standard Model	(2) Social Preferences Model	(3) Social Identity Model	(4) Standard Model	(5) Social Preferences Model	(6) Social Identity Model
Own payoff (β)	0.130***		0.346***	0.206***		0.714***
	(0.022)		[0.044]	(0.034)		[0.087]
Other's payoff when ahead (ρ)		0.469***			0.468***	
		(0.030)			(0.042)	
Other's payoff when behind (σ)		0.384***			0.246***	
		(0.037)			(0.067)	
Norms (γ)			2.688***			3.631***
			[0.236]			[0.451]
$0.1 \cdot \frac{\gamma}{\beta}$			0.777***			0.508***
			[0.067]			[0.027]
Observations	4290	4290	4290	2409	2409	2409
Log likelihood	-903.6	-900.5	-745.9	-484.1	-474.7	-420.5
AIC	1809	1805	1496	970.1	953.4	844.9
BIC	1816	1818	1509	975.9	965.0	856.5

Notes: 1. Standard errors (in parentheses) and bootstrapped error [in brackets] are adjusted for clustering at the individual level.
2. Significant at the *** 1 percent, ** 5 percent, and * 10 percent levels.
3. The number of observations in Panel A comes from 270 tax-framed and 154 neutrally-framed Democrats, each making 11 dictator choices for each of 3 endowments. The number of observations in Panel B comes from 130 tax-framed and 73 neutrally-framed Republicans, each making choices in the same situations.

We next examine the results of the neutral frame (columns 4 to 6). Both the AIC and BIC values for the *social identity model* (1595 and 1608 for Democrats, 844.9 and 856.5 for the Republicans, respectively) are smaller than those for the *standard* (2066 and 2072 for Democrats, and 970.1 and 975.9 for Republicans, respectively) or the *social preferences models* (1994 and 2007 for Democrats, and 953.4 and 965.0 for Republicans, respectively). A Vuong test of goodness-of-fit for each identity also shows that the *social identity model* fits better than the *standard* or the *social preferences models* for the neutrally-framed treatment ($p < 0.01$ for both Democrats and Republicans).

1.4.2.3 Parameter estimation

The conditional logistic regressions reported in Table 1.4 also give us parameter estimates of the tested models. The reported coefficients reflect the relative weight of each component in the utility function. Specifically, the coefficient for the dictator's own payoff, β, is an estimate of the average weight dictators place on their own monetary payoff. The coefficients for the other's payoff, ρ and σ, give estimates for how much the dictators care about the receivers' payoffs on average when they have higher (ρ) or lower (σ) monetary payoff than the receivers. The coefficient for norm ratings, γ, provides an estimate of the average weight subjects place on social appropriateness.[27]

For the *social identity model*, we can take advantage of the estimation structure and use the ratio of γ and β to estimate how much money an individual is willing to give up for a one unit increase in the norm rating. This can be seen as the equivalent of choosing an action that is deemed very socially appropriate over an action that

[27]Because the average norm ratings are a measured quantity which may include sampling error, we use bootstrapped standard errors for the *social identity model*. To construct the bootstrapped standard errors, we conduct 1000 replications. In each replication, we resample (with replacement) from the appropriateness ratings data (generated from the *norm elicitation experiment*) and construct an average norm function $N(\cdot)$. We then re-estimate the choice model based on the sampled norm function. The distribution of the coefficients across replications generates the standard errors.

is neutral. Allowing only changes in the monetary payoffs and norm ratings for the actions, we obtain:

$$dk_i/(dN(k_i|I_i, e)) = (\partial P_i/\partial N)/(\partial P_i/\partial k_i) = \gamma_i/\beta_i \tag{1.3}$$

We then multiply this ratio by 0.1 to get the dollar value of the trade-off, since each token in our experiment is worth \$0.10.[28]

We report the results of the *standard model* for our tax-framed treatments in column 1 of Table 1.4. For both Democrats ($\beta = 0.112$, $p < 0.01$) and Republicans ($\beta = 0.130$, $p < 0.01$), we find that the coefficient on the monetary payoff is positive and significant. That is, subjects are more likely to choose an action that has higher payoffs.

Next, we report the results of the *social preferences model* for the tax-framed treatments in column 2. We find that the dictators in our experiment care positively about the receivers' payoffs both when they make more or less money than the receivers. When the dictators make more than the receivers, we find for both Democrats ($\rho = 0.575$, $p < 0.01$) and Republicans ($\rho = 0.469$, $p < 0.01$) that the estimated ρ parameter is positive and statistically different from 0. Similarly, when the dictators make less than the receivers, we find for both Democrats ($\sigma = 0.234$, $p < 0.01$) and Republicans ($\sigma = 0.384$, $p < 0.01$) that the estimated σ parameter is positive and statistically different from 0. These estimates indicate that, under the *social preferences model*, our subjects exhibit social welfare preferences, a finding that reflects the overall finding in *Charness and Rabin* (2002).

We next report the results of the *social identity model* in column 3. Here, we find the coefficient on the payoff is positive and statistically significant for both identities ($\beta = 0.419$, $p < 0.01$ for Democrats and $\beta = 0.346$, $p < 0.01$ for Republicans). Further, we find that the coefficient on the tax-framed norm ratings is also positive

[28]Similar analyses using these ratios are also reported in *Davies et al.* (2001) and *Boskin* (1974).

and significant ($\gamma = 2.833$, $p < 0.01$ for Democrats and $\gamma = 2.688$, $p < 0.01$ for Republicans). The latter result suggests that subjects are more likely to choose actions associated with higher norm ratings.

For both Democrats and Republicans, we find that the magnitude of the coefficient on tax-framed norm ratings (γ) is larger than that on a subject's monetary payoff (β). That is, a subject's concern for the social appropriateness of an action outweighs her concern about the payoff of that action. Calculating $0.1 \cdot \gamma/\beta$, we see that tax-framed Democrats are willing to sacrifice \$0.68 for a one unit increase in appropriateness, while tax-framed Republicans are willing to sacrifice \$0.78 for the same increase in the appropriateness level.[29]

We next present the results of the *standard* and *social preferences models* for our neutrally-framed treatments in columns 4 and 5 of Table 1.4. Our results show that under the *standard model* (column 4), both Democrats and Republicans are more likely to choose actions that result in a higher payoff ($\beta = 0.191$, $p < 0.01$ for Democrats; $\beta = 0.206$, $p < 0.01$ for Republicans). For the *social preferences model* (column 5), dictators care about the receivers' payoffs both when they make more money ($\rho = 0.502$, $p < 0.01$ for Democrats; $\rho = 0.468$, $p < 0.01$ for Republicans) and when they make less money ($\sigma = 0.194$, $p < 0.01$ for Democrats; $\sigma = 0.246$, $p < 0.01$ for Republicans). In addition, comparing the coefficients for each model between treatments shows significant differences in some cases. For the *standard model*, Wald tests show that β differs significantly for Democrats ($\chi^2 = 8.32$, $p < 0.01$) and marginally significantly for Republicans ($\chi^2 = 3.47$, $p = 0.063$). For the *social preferences model*, there are marginally significant differences in ρ for Democrats ($\chi^2 = 3.73$, $p = 0.053$) and in σ for Republicans ($\chi^2 = 3.35$, $p = 0.067$).

[29]For the regressions regarding the *social identity model*, we check for multicollinearity by calculating the variance inflation factor (VIF). For Democrats, VIF = 3.35 for the neutral frame and VIF = 1.28 for the tax frame. For Republicans, VIF = 2.18 for the neutral frame and VIF = 1.10 for the tax frame. Though this seems to indicate low levels of multicollinearity (VIF values above 10 are typically considered a problem), we note that the interpretations of γ/β might still be affected.

Under the *social identity model* in the neutral frame (column 6), dictators are more likely to choose actions that lead to higher payoffs ($\beta = 1.043$ and 0.714, $p < 0.01$ for Democrats and Republicans, respectively). In addition, we find that individuals place more weight on the appropriateness of an action ($\gamma = 4.602$ and 3.631, $p < 0.01$ for Democrats and Republicans, respectively). Calculating $0.1 \cdot \gamma/\beta$ as above, we find that Democrats and Republicans are willing to pay approximately \$0.44 and \$0.51, respectively, for a marginal improvement in the appropriateness of their actions.

While the ratio of γ and β estimates how much money an individual is willing to give up to be norm compliant, the change in this ratio in the framed and neutral treatments expresses the impact of the persuasive influence of the frame on choice. We find that for both Democrats and Republicans, there is approximately a 53% increase in this ratio from the neutral to the tax frame, indicating a similarly large effect of frames on choices regardless of political affiliation.

1.4.2.4 Identity-dependent norms improve out-of-sample prediction

As a second test of the *social identity model's* ability to capture how frames alter norms and choice, we examine its performance in out-of-sample prediction. To do so, we first calibrate the model using only a fraction of our data. We then validate the model using the derived estimates to predict behavior in the remaining data. Specifically, we first randomly choose 30% of the subjects in the *choice experiment* to be in the validation sample. We perform out-of-sample forecasting by estimating the models' parameters using the choices of the remaining 70% of our subjects in the *choice experiment*. We refer to this as the calibration sample. We then use those parameters to predict the choices of the subjects in the validation sample.

In Figures 1.6, 1.7 and 1.8, we plot the actual vs. predicted behavior of dictators under the *standard*, *social preferences*, and *social identity models*, respectively. The top row shows the results for an initial endowment of 0, the middle row shows the

results for an initial endowment of 5, and the bottom row shows the results for an initial endowment of 10. The first two columns show the results for Democrats and the last two columns show the results for Republicans. The histograms represent the validation sample and the dashed lines represent the predicted sample.

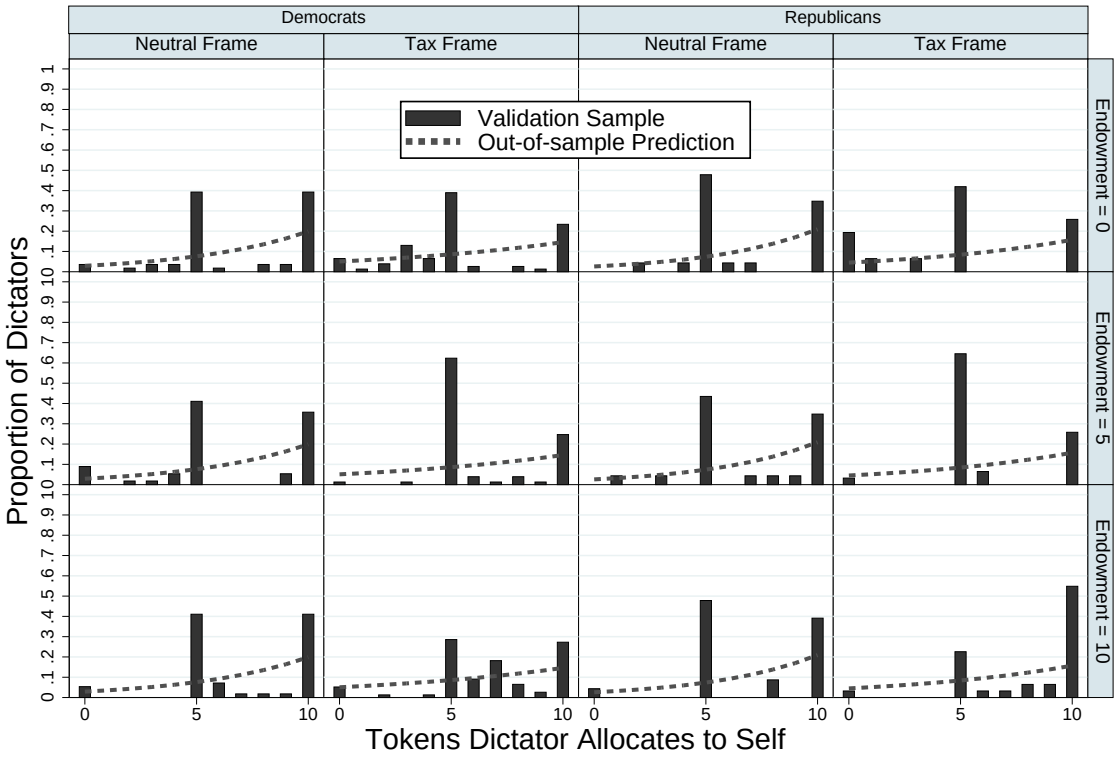

Figure 1.6: Actual behavior (validation sample, from 30% of all data) and out-of-sample predictions using the *standard model*.

As with the conditional logistic regressions, Figures 1.6-1.8 visually depict the better fit of the *social identity model* to actual behavior. Using the *standard model*, we see that dictators are predicted to be more likely to choose actions with higher payoffs and are most likely to keep all 10 tokens. Under the *social preferences model*, Democratic dictators are predicted to choose an equal split most often while Republican dictators are predicted to keep all 10 tokens most often. However, the *social identity model* correctly predicts that both keeping all ten tokens and an equal split

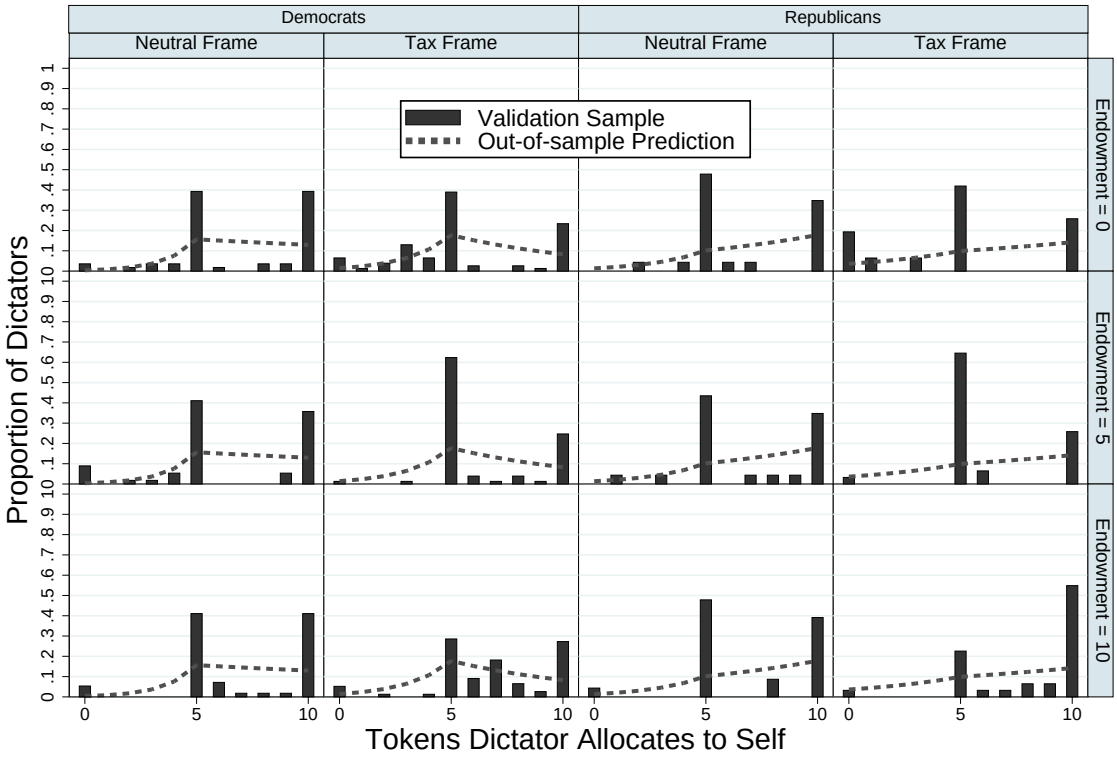

Figure 1.7: Actual behavior (validation sample, from 30% of all data) and out-of-sample predictions using the *social preferences model*.

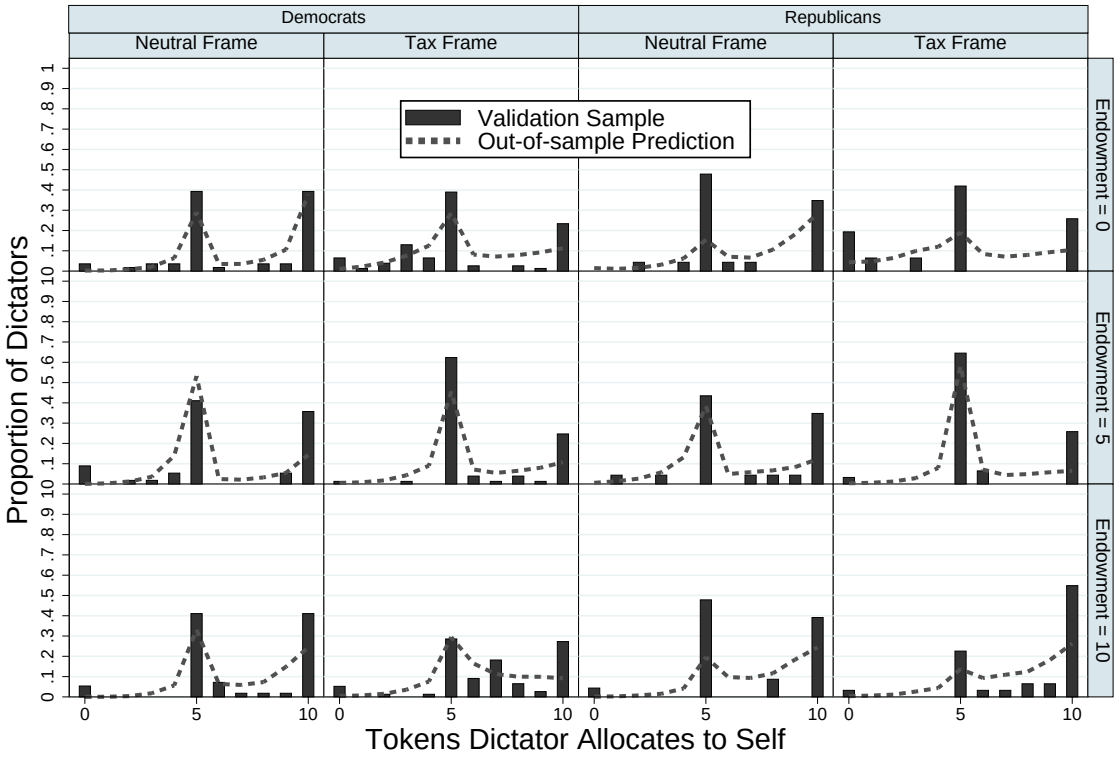

Figure 1.8: Actual behavior (validation sample, from 30% of all data) and out-of-sample predictions using the *social identity model*.

of the tokens are likely.

In Table 1.5, we present the quadratic scores of the different models when applied to our validation sample. Since the models' predictions are in the form of probability weights on the possible actions, we use the quadratic scoring rule to compare the performance of the models. When a subject in the sample takes action k, the quadratic score for a model that predicts the subject will have taken actions $(0, 1, 2, \ldots, 10)$ with probability $\boldsymbol{p} = (p_0, p_1, p_2, \ldots, p_{10})$ is:

$$Q(\boldsymbol{p}, k) = 1 - \sum_{i=0}^{10} (\delta_{ik} - p_i)^2$$

where $\delta_{ik} = 1$ if $i = k$ and $\delta_{ik} = 0$ otherwise.

Signed-ranks tests confirm the the *social identity model* yields higher quadratic scores than the *standard* and *social preferences models* in all cases ($p < 0.01$ for all comparisons).[30] This confirms that the *social identity model* predicts actual behavior more accurately than either the *standard model* or the *social preferences model*.

Table 1.5: Quadratic scores for the validation sample

Identity	Frame	Standard Model	Social Preferences Model	Social Identity Model
Democrats	Neutral	0.127	0.129	0.271
	Tax	0.103	0.137	0.240
Republicans	Neutral	0.125	0.131	0.225
	Tax	0.112	0.113	0.272

[30]These tests also show that the *social preferences model* yields higher quadratic scores than the *standard model* ($p < 0.01$) in all but one case. For Republicans in the neutral frame, this test shows that the quadratic scores between the *standard* and *social preferences models* are not significantly different ($p = 0.13$).

1.5 Discussion

The argument we make is that frames invite different interpretations of acts and outcomes because they evoke different norms. Our tax and neutral frame treatments follow *Kahneman* (2000) in that the experimental manipulation changes the description of the situation. We also vary the initial endowment within subject and show that variation in initial endowments affects behavior. As noted in the literature review, this latter result contrasts with the majority of prior work examining the effect of the initial endowment on dictator behavior.

Our analysis, presented in A.2, suggests that the significant behavioral change observed when we vary the initial endowment, is due to our within-subjects design. The interpretation we offer is that changes encountered by a subject in initial endowments act as a (procedural) frame (*Larrick and Blount*, 1997; *Kahneman*, 2000). This is consistent with the insights upon which Prospect Theory was built: changes from an initial point are salient.

This interpretation offers a way to harmonize the sometimes conflicting results of prior work on initial endowments in dictator games (e.g. *Krupka and Weber* (2013a) vs. *Dreber et al.* (2013)). Our data do not dispel the possibility that a single initial endowment, such as ($5, $5), brings with it norms (such as "do not take") that affect behavior (as in *Grossman and Eckel* (2015) or *Krupka and Weber* (2013a)). However, our results suggest that (procedural) framing effects emerge more strongly in within-subjects designs and with endowment distributions that are not "extreme" (all wealth to recipient or all wealth to dictator).

In addition, we find evidence that the impact of our tax and neutral frames emerge more strongly when there is more ambiguity about the social norms (*Dreber et al.*, 2013). In our games, we find evidence that Republicans report norms with greater variance (relative to Democrats) in the neutral frame ($p < 0.05$), which we interpret as greater ambiguity about the norm. We also find that when we impose the tax frame,

then it is the Republicans whose behavior is more affected by the frame ($p < 0.05$ for the interaction effect in identity, frame, and endowment). This suggests that greater norm ambiguity in the neutral frame makes the impact of the tax frame more significant for Republicans. The tax frame connects with a prescription or normative imperative for that identity. Though our results are consistent with the suggestion by *Dreber et al.* (2013), more would need to be done to explore the role of norm ambiguity and frame effectiveness.[31]

1.6 Conclusion

Prior research shows that rhetorical framing impacts a decision maker's conception of the acts, outcomes and associated contingencies for a particular decision. In this study, we provide insight into how framing works through two experiments. Using the context of Democratic and Republican identities in the U.S., we first find that framing affects norms, which in turn impact behavior. We also find that a model of social identity provides a tractable explanation for this effect. Specifically, we find that the identity model yields lower AICs and BICs in the conditional logistic regressions when compared to an outcome-based model, as well as better out-of-sample prediction accuracy.

Our study makes several important contributions to the literature. The first contribution is an improved understanding of framing. Though framing has a well-documented effect on behavior, we do not really understand why it works. This paper presents one mechanism for how framing impacts choice: frames evoke norms, which in turn influence choice. This offers one mechanism by which unstable preferences will be impacted by a frame.

Our study presents a novel method that allows sharper predictions for the likely impact of frames on behavior. That is, we directly measure the effect of frames on

[31] We thank an anonymous referee for pointing us in this direction.

norms. Previous research on how framing affects behavior often relies on more general intuitions, such as "we dislike losses."By contrast, we show that there are interactions between idiosyncratic characteristics (such as a person's social identity) and a frame that can be anticipated. This insight allows us to formulate a richer model of how frames affect decision making.

Our second contribution is to advance how we study social identity (*Akerlof and Kranton*, 2000; *Chen and Li*, 2009). Despite the central role of norms in identity-based choice models, previous work often relies on assumptions about these norms (see *Roy*, 1952; *Benjamin et al.*, 2010).[32] As a consequence, choice data alone cannot separately identify identity-dependent norms and behavior as the observed choice is a consequence of both an individual's utility over outcomes as well as her utility derived from norm compliance.[33]

By contrast, in our study, we separately and independently identify identity-dependent norms, thus overcoming several challenges associated with work on social identity. Our approach makes it possible to construct tests of the social identity model for those identities or situations where we do not have *ex-ante* strong intuitions regarding the norms. It also allows us to make specific predictions about the behavior we expect.

The broader implications of this study for policy makers regarding the use of framing language are both intuitive and striking: we may not be as politically divided as we appear to be. Two examples provide stark evidence that frames make us seem more divided than we are. At the time that the Affordable Care Act was receiving wide news coverage, Democratic strategists noted that re-naming the ACA

[32]These assumptions may have been necessitated by the fact that such identities are fluid, multiple, and socially-constructed (*Turkle*, 1997; *Shih et al.*, 1999). Also, research has shown that norms can vary from situation to situation (*Krupka and Weber*, 2013a; *Bicchieri*, 2005).

[33]*Charness et al.* (2014) identify the trade-off between identity and potential monetary considerations by varying whether people participate in a group activity as well as the size of the endowment received. However, their focus is on which sense of identity becomes salient in which circumstance rather than on how identity shapes norms.

to Obamacare would have a polarizing effect: "When the GOP turned the ACA into Obamacare they turned a bill that many GOP voters would like because it provided them affordable health care into a referendum on a president whom their voters hated" (*Pathe*, 2017). A similar impact of framing is well documented with respect to climate change. As *Leiserowitz et al.* (2014) note in their report from the Yale project on climate change communication, "...global warming and climate change are often not synonymous—they mean different things to different people—and activate different sets of beliefs, feelings, and behaviors, as well as different degrees of urgency about the need to respond." Our work offers both a path forward to study and predict the impact of framing on choice, but also an identity-based approach to ameliorate its negative effects.

CHAPTER II

Engineering Information Disclosure: Norm Shaping Designs

2.1 Introduction

Interface design in social media systems can greatly influence when, how much, and why people disclose private information. For example, the addition of a user interface element that displays friends' birthdays (Figure 2.1) may lead a user to perceive that sharing birthday information is the norm and subsequently to share this data with the system. However, the mechanisms through which interfaces affect disclosure behavior are still poorly understood. In particular, we understand little about how interfaces signal social norms and, in doing so, encourage information disclosures. We broadly know that signals of norms can assist newcomers and both encourage and discourage pro- or anti-social behavior. We also broadly understand that decision contexts may be altered to nudge behavior towards target outcomes. However, we know much less about how norms, contexts, and interfaces combine to form social guideposts that shape users' beliefs about what kind, and how much, information is acceptable to reveal. Work in economics and psychology describes a social learning mechanism through which actors observe behavior that informs their beliefs about what is appropriate to do in a particular context and then change their

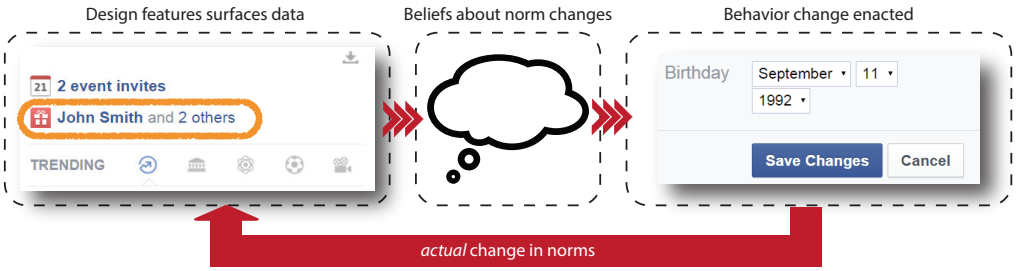

Figure 2.1: How designers influence information sharing.

own behavior to conform with those (newly updated) social expectations. In this paper, we design an experiment that tests a causal pathway from design choices, via changes in beliefs about what is appropriate, to subsequent behavior change.

Shaping behaviors through user interface design is a well-established principle in modern Human-Computer Interaction (HCI) practice. Interfaces often provide cues or affordances to help a user (or communities *Kraut et al.* (2012)) achieve some end-goal, and many designs that are successful are encoded as patterns (using the terminology of the software engineering literature). For example, a multi-layer pattern hides complexity so that the end-user is not overwhelmed *Adar et al.* (2013), and a voting pattern uses a thumbs-up or down icon to encourage voting *Crumlish and Malone* (2009).

In contrast to these helpful patterns, there are other patterns that are not intended to benefit the user directly, but rather to serve the interest of another party–such as pushing users to provide personal information, or even "dark" patterns that nudge users to take security risks *Conti and Sobiesk* (2010). As an example, forced disclosure patterns require users to fill in information before gaining access to a service. The preference that system owners and builders hold for certain designs over others may be motivated by explicit corporate interests (e.g., advertising revenue or engagement) and behavior targets (e.g., design A drives more photo sharing than design B). These pressures might mean that incentives to maintain privacy conflict with systems' goals

to reduce privacy-maintaining actions and may implicitly create an impression of information sharing norms *Young and Jordan* (2013).

Norm-shaping patterns need not be positive or negative. However, the fact that such patterns can be used both ways makes them particularly insidious, because users cannot infer what the designer's intent was or the downstream consequences of their behavior may be *Debatin et al.* (2009). Policy makers and well-intentioned designers have no mechanism for assessing how their design choices shape norms. Our claim is that *design choices have the power to engineer personal information give-away by changing beliefs and, subsequently, behavior.* The implications are not only that the user has revealed more information to the system than they may have intended, but that this opens up the downstream uses of this data in ways that the user never envisioned (e.g., *Acquisti and Gross* (2009)).

In this paper we experimentally identify the impact of design on the perception of appropriate behavior and we test the impact of norm-shaping design patterns on subsequent information divulging behavior. We find that subjects' perception of what images they personally would find acceptable to post on a social media site can be altered when we manipulate the context to contain more "risky" or "explicit" images. Further, we find that a shift in perception affects subsequent information divulging behavior in a subsequent task as well as the advice given to others about revealing personal information.

Our main contribution is to experimentally unpack and identify a causal pathway by which design patterns can work to affect disclosure: they can modify perceptions of appropriate behavior which, in turn, impact subsequent behavior. We show a chain of influence, where designers may nudge users' beliefs about what is appropriate (in a situation), and subsequently alter users' behaviors in that situation. Our second contribution is to demonstrate that the shift in perceptions can leave a larger footprint on user behavior than one might think. This is because the shift in perceptions of

appropriate behavior also impacts the advice that a user gives others. This gives rise to a cycle of nudging individual behavior through design patterns. The cycle begins with a design pattern that shapes *perceptions* of what is personally acceptable to do, which then nudges behavior which then shapes the norms of the community through altered behavior and through altered advice given to others (Figure 2.1). Further, we demonstrate that nudges in a one medium (images) affect norm perceptions for that medium (posting images on a social network website) and have a spillover to a second domain (revealing information to the experimenter) and advice given to another user (e.g., advice on what is appropriate to reveal). Our work leverages a novel experimental design, which combines methodologies and theory from experimental economics with HCI methods. Taken together, our findings have significant implications for security and privacy.

2.2 Related Work

Much of the modeling of information disclosure decisions rests on the assumption that a user can at least tell you how much she values privacy. Various studies investigate how people navigate the trade-off between sharing private information and other instrumental goals such as better recommendations or financial remuneration *Acquisti* (2004). A subset of these studies highlight the dichotomy between professed privacy attitudes and actual self-revelatory behavior *Acquisti and Gross* (2006); *Acquisti and Grossklags* (2005); *Spiekermann et al.* (2001); *Tedeschi* (2012).

Some of the (numerous, and not mutually exclusive) explanations for the dichotomy reside in the hurdles that hamper individuals' privacy-sensitive decision making. In particular, this literature posits that the disparity between professed attitudes and behavior stems from either uncertainty due to incomplete information about one's preferences or from uncertainty about which social rules apply in that context *Acquisti et al.* (2015). Users may experience *preference uncertainty*–that is,

they have a vague sense of, or they just don't know, how much they value privacy. Alternatively users know their preferences but are trying to figure out how to trade-off between their preferences for privacy and the social norms and expectations that others have of them to share information.

These hurdles make privacy *attitudes* appear inconsistent and/or easily malleable. The hurdles make disclosure *behavior* highly contextually sensitive and surprising or seemingly contradictory. Both explanations (uncertainty over preferences or uncertainty over social norms) imply that users may rely on social cues (such as the behavior of others) as they decide how much to disclose *Tsai et al.* (2011). As an example, Acquisti et al. *Acquisti et al.* (2012) find that disclosure behavior is comparative in nature: People's willingness to divulge sensitive information depends on judgments about others' readiness to divulge that information.

This evidence is suggestive of an interplay between others' behaviors, perceptions of social norms, and one's own behavior. And although sharing behavior, as well as the drivers of personal disclosures, have been investigated from a number of different disciplinary angles *Reinmuth and Geurts* (1975); *Tourangeau and Yan* (2007), several new questions emerge. How critical are social and contextual cues to shaping the *perception* of norms as well as the nudging of individual behavior? Do context and social cues operate on behavior by changing beliefs about the norms? The implication of such relationships is a cycle of nudging individual behavior through design patterns which, in turn, affects the *perceptions* of norms, which then nudges behavior which then shapes the norms of the community.

The importance of norm-shaping in social media and computer-supported cooperative work has been recognized by a number of researchers (many summarized in *Kraut et al.* (2012)). As social media systems gain popularity, designers of social media interfaces are creating a set of interface heuristics (e.g., having a "conversational" style question such as "What are you up to?" instead of "Type status here." to make

users more comfortable) *Crumlish and Malone* (2009). These design patterns have evolved as a result of conventional wisdom, aesthetic concerns, and ad-hoc experimentation, but rarely through principled studies and even more rarely in information disclosure contexts (though see *Wang et al.* (2014, 2013)). However, much of this work focuses on the impact of these signals on the end-outcome (behavior) and does not focus on why the behavior changed in response to the social signal. One pathway by which a social signal (such as observing what others are doing) can impact behavior is by changing beliefs about what is appropriate to do or the social norms.

A long tradition of work in psychology and, later in economics, shows that by observing others, people learn what behaviors are considered appropriate as well as expected. This work predicts a positive relationship between one's action and what one observes others doing *Bardsley and Sausgruber* (2005); *Cialdini et al.* (1990). In psychology, the classic experiments showing this influence involve observing how an individual's judgment of the length of a line segment varies depending on the responses of others *Asch* (1956); *Deutsch and Gerard* (1955). Recent work in economics has predominantly demonstrated this relationship in public goods games *Bardsley and Sausgruber* (2005); *Gächter and Renner* (2003); *Shang and Croson* (2009).

Work in human-computer interaction and computer-supported cooperative work has sought to understand the drivers behind information disclosure such as "folk models" (*Wash*, 2010), privacy concerns (e.g. *Stutzman et al.*, 2012) and preference models (e.g., *Knijnenburg and Kobsa* (2013); *Knijnenburg et al.* (2013)). Considerable work in psych (e.g. *Cialdini et al.* (1990); *Deutsch and Gerard* (1955)), economics (e.g. *Bandiera et al.*, 2005; *Krupka and Weber*, 2009) and applied design (e.g. *Harper et al.*, 2010; *Sukumaran et al.*, 2011) demonstrated that showing a user what others are doing (or think should be done), leads to conformity in action (or with respect to normative expectations). However, this work does not establish casual mechanisms by which this correlation is observed.

More recently, researchers have begun to unpack the pathway by which this influence occurs: observing others influences the actor's beliefs about what actions are appropriate which in turn affects behavior *Krupka et al.* (2015). Both economics and psychology offer a rich empirical body of work demonstrating the impact of others' behavior on an actor's norm compliant behavior. However, the pathway from observing others, to changing one's normative beliefs, to changing one's behavior in the context of information disclosure is not as well-understood (though the fact that such a pathway *exists* has been demonstrated in *Acquisti et al.* (2012)).

A great deal of existing research on privacy in the context of social media platforms such as Facebook relies on self-reports (e.g. *boyd and Hargittai*, 2010; *Stutzman and Kramer-Duffield*, 2010; *Stutzman et al.*, 2012). While informative, it is difficult to directly assess how a design element may impact perceptions and behavior. In our research we use experiments to test the impact of design patterns on users' perceptions of information sharing norms. Specifically, we manipulate the types of information our target users see others providing and use that to test the pathway from observation to perceptions of social norms. To accomplish this we adapt instruments developed in previous work by Krupka and Weber *Krupka and Weber* (2013b) and embed them in an experiment. Our primary goal is to demonstrate that perceptions about norms that govern information disclosure can be manipulated and affect subsequent disclosure behavior. Specifically, we establish a causal pathway from beliefs about appropriate sharing, to disclosure, and to advice to others about disclosure.

2.3 Experiment

2.3.1 Overview of Design

We would like our experimental data to accomplish three goals. The first is to directly identify perceptions of norms associated with divulging information. Second,

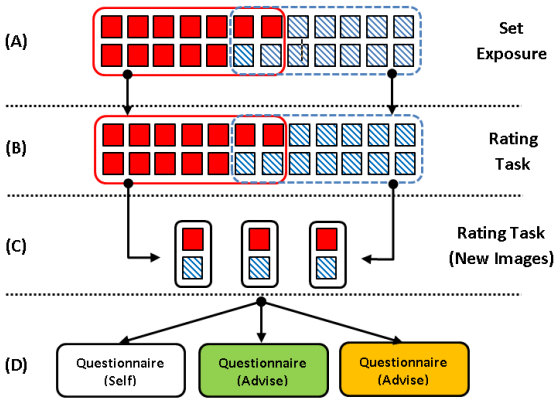

Figure 2.2: Subjects were randomly assigned into either the R condition and saw the R images (solid red squares) or the PG condition and saw the PG images (dashed blue squares). In addition, subjects also saw two R and two PG images (overlap in the middle). During *Set exposure* (A), participants saw their assigned images in a group. In the *Rating Task*, they rated each of these images individually (B). In the *New Image Rating Task* (C), they rated one of three possible sets of new R and PG images. Finally, subjects were randomly assigned to one of three possible questionnaire condition (D).

the data should allow us to test whether beliefs about norms can be nudged with norm-shaping design patterns, and third, whether subsequent behavior is affected. To accomplish these goals, we designed an experiment that tests how observing others' behaviors can influence (1) personal beliefs about information disclosure, (2) subsequent disclosure of information about oneself, and (3) subsequent advice to others about appropriate disclosure of information about themselves.

At a high level, our experiment tests whether people exposed to more (or less) provocative images display different perceptions of what is acceptable to share and subsequently change their information sharing behaviors. Our high level hypotheses are that more provocative imagery shapes norm perception and nudges subsequent behavior. We chose posting images (or "selfies") as our social media context because we were able to achieve a high degree of experimental control through systematic variation in the images shown to subjects (see Figure 2.3). More broadly though,

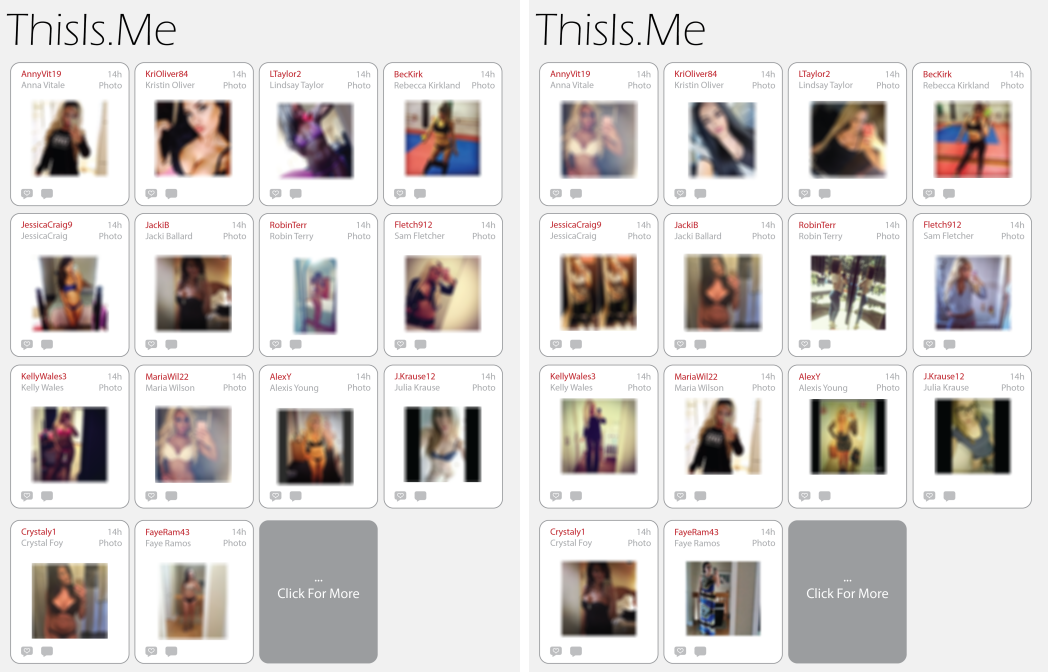

Figure 2.3: Examples of the ThisIs.Me page that subjects saw during *Set Exposure*. R individuals saw the image on the left while PG individuals saw the image on the right. The images in the experiment itself were not blurred.

designers have control over how and where widgets are shown; our experiment focuses on one such possible manipulation in one possible context. Though the experiment itself is highly controlled, it resembles what might actually be seen in real designs (we further explore generalizabiltiy in the discussion).

The experimental design consists of four steps. The first step exposes subjects to different image-exposure sets. They are randomly assigned to see images that are either more or less risky/provocative (in Figure 2.2A the solid red tiles depict more provocative images and the dashed blue tiles depict less provocative images). Throughout our analysis we will refer to the more provocative images as "R" and the less provocative images as "PG." In study 2, described below, we obtain independent ratings on each picture to be able to categorize them on appropriateness and on attractiveness.

In the second step, we ask subjects to rate how personally appropriate they find

each of the images they received in the image-exposure set (Figure 2.2B). In the third step, we show them 2 new images that were not originally part of the image-exposure set and ask them to rate how personally appropriate they find the new images to be. The two new images contain one R (solid red tile in Figure 2.2C) and one PG image (dashed blue tile Figure 2.2C). We created three pairs of new R/PG images for this step which are matched to be similarly appropriate (ratings from study 2, described below, are used to do this matching). However, subjects are only shown *one*, randomly selected, new R/PG pair. We will refer to these images as "new images" in our analysis.

In the fourth step, we observe subjects' disclosure behaviors and advice-giving behavior through their responses to a questionnaire. Subjects are randomly assigned to respond to one of three possible questionnaires. Either they respond to a questionnaire about themselves or they are asked to be an adviser to another fictitious person who either has a more "vanilla" or a more provocative, "cinnamon," personality.[1]

The design of the main experiment consists of a 2 (initial image exposure set is R or PG) $\times$ 3 (new image pair #1, #2, or #3) $\times$ 3 (questionnaire regarding self or advice to vanilla or advice to cinnamon person) design. Further, we minimize self-image concerns, where a subject may assess the relative physical physique between image and him/herself, by giving our male subjects female selfies and vice versa. Table 2.1 summarizes the treatment conditions in the main experiment. A subject could only be assigned to one of these cells. In what follows we describe each step in detail and relate it back to our high-level hypotheses.

2.3.2 Step one: initial set exposure

To create the environment for step one, we generated a fake photo-driven social media site, ThisIs.Me (see Figure 2.3), and told subjects that we were introducing

[1]We did not describe the fictitious person as vanilla or cinnamon to subjects, but adopt that language here for ease of description.

	Questionnaire Self	**Questionnaire Advise Cinnamon-type**	**Questionnaire Advise Vanilla-type**
R	R - Self	R - Advise (Cinnamon)	R - Advise (Vanilla)
PG	PG - Self	PG - Advise (Cinnamon)	PG - Advise (Vanilla)

Table 2.1: Participants were randomly assigned into one of these six conditions. Within each cell, they were randomly shown new image pair #1, #2, or #3. A participant was only assigned to one of the cells and did not participate in any of the other conditions. *Note: the text in the cell describes the exposure set and questionnaire type a subject received.*

a new feature for the site. Our instructions read: "This plugin would scan pictures that you choose to post. For certain images, the plugin would ask you to wait for 10 minutes before deciding to post an image and then, after 10 minutes, it would ask you:'Do you want to post this?"' We informed subjects that their job was to teach the plugin what kinds of pictures the subject might later regret or wish that they had not posted. The instructions explained that "To help the plugin learn how to advise you, we will give you a set of 'training' images that others have posted [...] and ask you to rate how appropriate you think they are." Subjects used a scale from 1 "very inappropriate to post" to 6 "very appropriate to post" to rate how *personally* appropriate they felt it was to post to our hypothetical social network site.

After reading the instructions, subjects were exposed to the initial set of images all at once on one screen–we call this the initial set exposure. The initial set exposure mimics a strategic surfacing of images to a new user that might happen when they first log onto the site. The intent is to shape the user's perceptions of what others are posting and consider acceptable. The question we test in step two is whether initial exposure to an R or PG set can also change what the user *personally* believes is acceptable for him or herself to post on our social network site.

To manipulate the type of selfie posting behavior subjects saw in the initial set exposure, we collected two types of images from existing Instagram accounts. We collected one R (where the individual is mostly undressed) and one PG image (where the individual is mostly dressed) from the same Instagram account so that our R and

PG sets contained comparable images. Pairs of images were selected so that the figure was roughly in the same pose and the picture was taken with the same camera angle. The initial R set contained a total of 14 images selected from Instagram accounts: 12 R images and also 2 "overlapping" PG images. The initial PG set also contained 14 images: 12 PG and also 2 "overlapping" R images.

We used the 2 PG and 2 R overlapping images to create a common group of 4 images that subjects in both the R and PG conditions saw. This is visually depicted in Figure 2.2A by the solid red and dashed blue outlines overlapping over the 4 central tiles. These overlapping images are used in the analysis because they allow us to test for how the initial set exposure affects subject perceptions of appropriateness on identical images.

2.3.3 Step two: Rating task

To test whether the initial exposure set impacts personal beliefs about what is appropriate to post, subjects rate each image from the initial set one at a time on subsequent screens (although the order was randomized at the subject level) (Figure 2.2B). Our participants used a Likert scale to rate how *personally* appropriate they felt it would be to post the image on the ThisIs.Me web site. The critical point here is that subjects were asked to tell us their *personal* opinions about how appropriate an image was to post. We purposefully did not ask them to tell us whether they thought *others* on the site would think the post was okay, but instead focused on how their personal views about posting the images were affected. Put another way, it would be very reasonable for an end-user to infer what others think is appropriate to do based on their posting behavior. What this design tests is how the behavior of others changes what the *subject* thinks is appropriate for *the subject herself* to do. With this data we can test our first hypothesis:

Hypothesis 1. *Individuals exposed to the R set will rate the pictures in the R set to be more personally appropriate than those exposed to the PG set.*

Once subjects finished rating the selfies from the initial set, they were randomly assigned to rate one of three pairs of new images (Figure 2.2C).[2] In this rating task, participants saw one new R selfie and one new PG selfie. We introduced these new images so that we can test the spill-over effects of exposure to the initial set onto new images.

Hypothesis 2. *Individuals exposed to the R set will rate new pictures to be more personally appropriate than those exposed to the PG set.*

2.3.4 Step three: Questionnaires

In step three we test whether subjects are more likely to divulge (or advise another to divulge) information in a subsequent task. To test whether the effect of initial set exposure extends beyond changing beliefs about what is appropriate to do *within* the social media site, our subjects filled out a questionnaire (23 questions in total). The questionnaire immediately followed steps one and two, but we randomized subjects into one of three conditions.

In the self-questionnaire condition, subjects were told that for the ThisIs.Me site "...users may create a profile card that is visible to other members of the site. The profile card will have your username, your selfie and your answers to the series of questions in this section. If you prefer to not have an answer show up in your profile card, you may choose to skip that question by selecting the 'skip' option." The questionnaire (a complete copy is available in the supplementary materials) contained

[2]To control for the possibility that the difference in attractiveness of the individual in the selfies may affect subjects' perception of appropriateness, we provided three different sets of new images, as opposed to just one. Each image was rated on attractiveness in study 2, described below, and all analyses control for this.

items that rang from less intrusive (e.g. "How often do you hold the door open for someone?") to very intrusive (e.g. "Did you ever have sex with someone who was too drunk to know what they were doing?"). These questions were scaled on intrusiveness in previous work by Acquisti et al. *Acquisti et al.* (2012).

We also used the responses from the self-questionnaire to create two types of potential new users to the site. We selected one set of responses that were more conventional–we termed this our vanilla new user. In another case we selected a set of responses that were not conventional –we termed this our cinnamon new user. We chose these two types of respondents to mimic a scenario where a new user's answers are well outside of how most others are answering the questions.

In the advise-questionnaire conditions, subjects saw the new user's responses to the questionnaire items and subjects were asked to provide advice to the new user on whether to include this on the profile card. Subjects saw either the vanilla or the cinnamon new user's responses and were instructed: "A user of this site has filled out this questionnaire, but has not yet submitted and published their answers on their profile card. Given their potential answers, help them decide which questions they should choose to submit and publish and which they should choose to 'skip."'

Because subjects were randomized into the three questionnaire conditions (self, advise-vanilla and advise-cinnamon), we can test three hypotheses. We can use the self-questionnaire responses to test how initial set exposure affects the likelihood that a subject chooses not to divulge information about him/herself by skipping some questions. Second, we can test whether initial set exposure affects the advice a subject will give to a cinnamon or vanilla set of responses.

Hypothesis 3. *Individuals exposed to the initial R set will skip fewer questions in the self-questionnaire condition than those exposed to the initial PG set.*

Hypothesis 4. *Individuals exposed to the initial R set will advise a cinnamon type to skip fewer questions than individuals exposed to the initial PG set.*

Hypothesis 5. *Individuals exposed to the initial R set will advise a vanilla type to skip fewer questions than individuals exposed to the initial PG set.*

2.3.5 Study 2: Baseline ratings

Lastly, since the selfies may vary in appropriateness and attractiveness even within the R and PG groups, we ran a second study with different subjects to collect baseline appropriateness and attractiveness ratings for each of the images we used in the our main study. Subjects saw all of the R and PG images together and then rated the images. Thus, their appropriateness ratings were not made after being exposed to a biased set of R or PG images. All of our regressions control for the baseline appropriateness and attractiveness ratings. In our analysis we refer to these ratings as our "baseline appropriateness rating" and "attractiveness ratings." Together, we refer to them as our "controls."

2.4 Results

To pilot the experiment we utilized a pool of students at a large academic institution (undergraduate and graduate students). To achieve greater analytical power and to generalize beyond this pool we utilized Amazon's Mechanical Turk infrastructure. A total of 387 Turk workers participated in our study. Of those, 305 participated in our main study (105 female and 200 male) while 82 (38 female and 44 male) participated in study 2. It took subjects an average of 6-7 minutes in both studies. We restricted participation to Turkers who have had at least 10,000 approved HITs and a HIT approval of at least 98%. Subjects were paid $0.50 for completing the survey.

We note that our Turk-based results are consistent with our pilot experiment, giving us some confidence in the stability of the results across populations.

	(1) Images: Overlap R	(2) Images: Overlap R	(3) Images: Overlap PG	(4) Images: Overlap PG
Dependent variable: Appropriateness rating				
R set exposure	0.328**	0.299**	0.286***	0.295***
	(0.145)	(0.125)	(0.0917)	(0.0891)
Baseline appropriateness rating		0.940***		0.713***
		(0.335)		(0.195)
Attractiveness rating		0.518		-0.200
		(0.597)		(0.187)
Baseline rating differences		0.227		0.250***
		(0.248)		(0.0735)
Attractiveness rating differences		(Omitted)		(Omitted)
Male		(Omitted)		(Omitted)
		()		()
Constant	2.563***	-2.723	5.009***	1.563
	(0.0972)	(2.129)	(0.0634)	(1.653)
Observations	610	610	610	610
R-squared	0.015	0.265	0.021	0.162
Subjects	305	305	305	305

Robust standard errors in parentheses, clustered on ID.
*** p<0.01, ** p<0.05, * p<0.10

Table 2.2: OLS regressions comparing appropriateness ratings on the 4 overlapping images by initial R or PG set exposure.

2.4.1 Rating task

2.4.1.1 Initial set exposure affects image appropriateness ratings

To test the influence of the initial set exposure on personal appropriateness ratings, we regress the appropriateness ratings subjects provided on the 4 overlapping images depicted in Figure 2.2A, on a dummy variable for the initial set exposure (if "R" then "R set exposure" takes the value of 1 and 0 otherwise) and on controls. We find evidence consistent with Hypothesis 1 (see Table 2.2).

Looking at column (1) of Table 2.2, R-exposure set subjects rate the overlapping R images to be more appropriate to post than those exposed to the PG set ($p<0.05$). This effect does not diminish when we include controls for the baseline appropriateness ratings, the attractiveness ratings of the individuals depicted in the selfies, and the differences in baseline appropriateness ratings between the overlapping R and PG

images (column (2)) ($p<0.05$). We find a similar effect of set exposure on the ratings of overlapping PG images. Set exposure to R increases personal appropriateness ratings on the PG images (column (3), $p<0.01$) and is not affected by controls (column (4), $p<0.01$).

We also wish to test the impact of the initial exposure set on how subjects rate the new images, Hypothesis 2. To do so, we regress appropriateness ratings for the new images on a dummy variable for initial set exposure to R ("R set exposure") and find evidence that supports Hypothesis 2. We report the coefficients of this regression in (Table 2.3).[3]

We find that those initially exposed to the R set rate the new selfie to be more appropriate to post than those exposed to the PG set. This is true when rating the new R image (column (1), $p<0.01$) and the new PG image (column (3), $p<0.01$). These results are also not sensitive to including controls for the selfie and a dummy for the participant's gender (columns (2) and (4)), ($p<0.01$).

2.4.2 Questionnaire

2.4.2.1 Initial set exposure affects information disclosure behavior

Since the questions in our questionnaire vary by intrusiveness, we expect to see more individuals skipping the more intrusive questions relative to the less intrusive questions. However, we also hypothesize that those participants who were initially exposed to the R set would skip fewer questions (Hypothesis 3) relative to those initially exposed to the PG images.

Figure 2.4 graphs the frequency of skips in the self-questionnaire by initial set exposure (R or PG). The red (bottom) line plots the skip frequency of subjects who

[3]We note that solving for differences between the treatments using a generalized linear model is mathematically equivalent to ANOVA (ANOVA being a particular case of the linear regression model with factor levels represented by dummy variables–e.g. 1 for the R group and 0 for PG). Thus, the analysis and conclusions drawn from regression and ANOVA are equivalent here.

	(1) Image: New R	(2) Image: New R	(3) Image: New PG	(4) Image: New PG
Dependent variable: Appropriateness ratings				
R set exposure	0.523***	0.513***	0.290***	0.292***
	(0.149)	(0.147)	(0.0937)	(0.0895)
Baseline appropriateness rating		(Omitted)		1.322***
				(0.427)
Attractiveness rating		0.204		0.313
		(0.730)		(0.642)
Baseline rating differences		0.307		(Omitted)
		(0.718)		
Attractiveness differences		-0.407		0.363
		(1.457)		(0.890)
Male		(Omitted)		0.192
				(1.010)
Constant	2.082***	0.632	5.335***	-3.614
	(0.0975)	(2.919)	(0.0716)	(3.924)
Observations	305	305	305	305
R-squared	0.039	0.082	0.030	0.127
Subjects	305	305	305	305

Robust standard errors in parentheses, clustered on ID.
*** p<0.01, ** p<0.05, * p<0.10

Table 2.3: OLS regressions comparing appropriateness ratings on the 2 new images by initial R or PG set exposure.

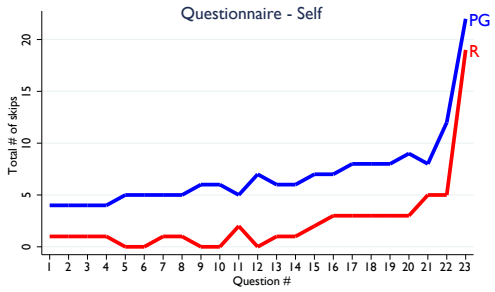

Figure 2.4: Frequency of skipped questions in the self-questionnaire condition by initial R (red line) and PG (blue line) set exposure.

were initially exposed to the R set and were answering the self-questionnaire. The blue (top) line plots the skip rate for those respondents who were initially exposed to the PG set.[4] Consistent with Hypothesis 3, for all questions, the skip rate for those exposed to the R set is lower than those exposed to the PG set.

To formally test whether the differences observed in Figure 2.4 are significant, we perform a logistic regression. We regress a dummy variable (1 if the question is skipped and 0 otherwise) on a dummy variable for whether the individual was initially

[4]The graphs plots the skipping frequency from least frequently skipped to most frequently skipped question on the x-axis.

exposed to the R set ("R set exposure"). Consistent with Hypothesis 3, individuals initially exposed to the R set are significantly less likely to skip questions (average discrete effect is 6.75% at p<0.10).

2.4.2.2 Initial set exposure affects advice about disclosure

Finally, we test the influence of the initial exposure set on advice-giving behavior. Recall, that subjects who received the advice-questionnaire, saw a fictitious vanilla or cinnamon subject's responses and were asked to advise whether each response should be skipped or posted as part of the profile. Figure 2.5 graphs the total number of advised skips by question (x-axis) and by cinnamon (left panel) or vanilla condition (right panel). The red line plots the total advised skips made by subjects who were initially exposed to the R set and the blue line plots the total advised skips made by subjects initially exposed to the PG set.

Looking just at the advice given to the cinnamon type (left panel), we see that initial exposure to the R set causes subjects to advise the cinnamon type user to skip fewer questions than initial exposure to the PG set. This difference is not visually apparent in the advice given to the vanilla type individual (right panel).

To formally test whether these differences are statistically significant, we regress a skip-dummy (which takes the value of 1 if the advice was to skip and 0 otherwise) on whether the adviser was initially exposed to the R (dummy is equal to 1) or PG set (dummy takes value of 0). Consistent with Hypothesis 4, exposure to the R set makes individuals less likely to advise the cinnamon type user to skip questions than exposure to the PG set (average discrete effect: 6.34%, p<0.10).

By contrast, we find no such differences when advice is given to the vanilla type of user. Initial exposure to the R or PG set, does not significantly affect the skipping advice given to a vanilla type user (p=0.406). Hypothesis 5 is not supported.

These results suggest that the effects of set exposure spill over into other areas of

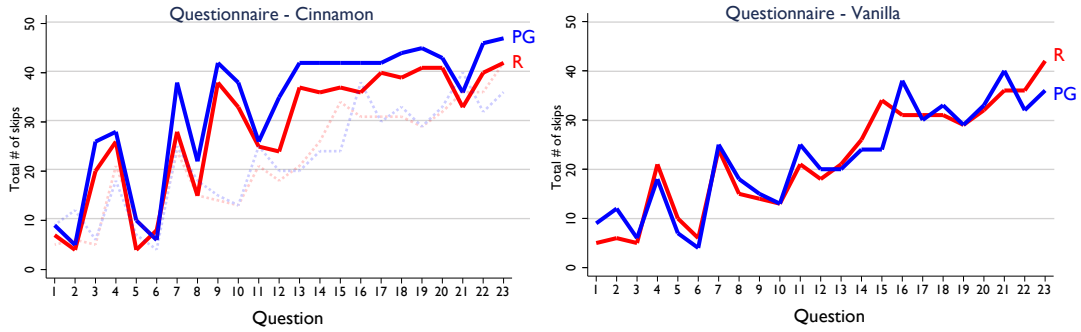

Figure 2.5: Total number of *advised* skips broken out by whether the adviser was initially expose to the R set of images or the PG set. The left panel plots the advice given to a cinnamon individual and the right panel plots the advice given to a vanilla individual. The lighter lines on the left panel are the superimposed lines from the right (vanilla) panel. This is done to ease comparison.

information disclosure. Further, they suggest that the effects of set exposure on advice about information disclosure impact advising when it is, in some sense, very desirable that they do so. If the goal is to get interesting and information rich responses posted, then set exposure is an effective tool.

2.5 Discussion

Given the ubiquity of social media systems there is a fundamental need to understand and mitigate against manipulative designs that lead to over-sharing. Users of these systems, many of whom are part of sensitive populations, are not always aware of the potential consequences of sharing their personal information (e.g., SSN inference, social phishing attacks). Further, as social media systems gain popularity, designers of social media interfaces are creating a set of interface heuristics *Crumlish and Malone* (2009).

While in some situations these heuristics emerge to satisfy certain design aesthetics, in many cases they are a result of conventional wisdom, theory, and scientific literature about motivating continued use of these systems and may favor designs

that increase sharing of private or information rich data. For example, consider the following UI components (some existing, some hypothetical):

1. A widget that indicates that 3 of your friends have a birthday today. This element shows you that your friends have shared their birthdays and may change your belief about the prescriptive norm of sharing birthday information. A change in beliefs about how acceptable or frequent others' sharing is, may increase your own propensity to share.

2. A quiz element indicating "90% of your friends took the salary quiz, find out how you rank." This element explicitly indicates how many of your friends took the survey, again encouraging a change in your belief about norms around sharing financial information that may result in your sharing such information.

3. A feed element displaying a recent photo album posted by a friend with 3 representative images. Such an element surfaces a friend's behavior and encourages the belief that posting photo albums is a norm. More subtly, as we have shown, the strategic choice of representative images can also distort an end-user's models of the type of pictures that should be shared. For example, the album may contain 90% pictures of sunsets, by which a random selection of representative images would likely show 3 sunset images. However, algorithms could surface non-representative images that contain flesh tones, have comments, or tagged friends. The strategic selection of such images can change behavior (posting more swimsuit pictures, less sunset pictures) by changing perceptions of how acceptable certain images are.

4. A site ad for an application that "helps you track your sexual contacts." This ad's request for highly sensitive information is intentionally targeted to be far outside the norms of typical requests; it acts to erode taboos and tells users something about the norms (i.e., what is acceptable to talk about). Through

the mere act of asking, designers give users the impression that acceptable topics for sharing encompass more provocative topics. Further, such a request could be used as part of a two-steps-forward-one-back strategy that explicitly asks end-users to share highly sensitive information so that other less sensitive, related requests–tracking your dating history–are perceived as more acceptable to share. Through this strategy designers can affect people's perceptions of norms regarding the acceptability of sharing the less sensitive information (after all, it is not sexual contacts!) and increasing the likelihood they do so.

Our results can be used to both safeguard and educate users, designers, and legislators. When automated, instruments can be developed to test new interface changes and provide early warning to users and those developing legal frameworks to better monitor, understand, and counteract the inappropriate use of these patterns and to enhance security.

Our research impacts the broader discourse and development of tools for the study of user interfaces as embodiments of social norms and other aspects of the culture and organization that the interface represents. Further, the combination of survey research and experimental economics, that allows us to explore differences between preferences and behavior and to explore issues of interface design, is a model that may be exploited in analogous research.

However, there are limitations to our findings that are important to mention here. Because of our experience with the more controlled pilot pool, we believe that the MTurk population is suitable for this experiment. Further, we purposefully chose a situation (posting pictures/questionnaire) that captures common online activities (from dating sites to Snapchat to Pinterest) where this type of information is routinely divulged. However, we believe it is also worth expanding the experiment to work within existing social media frameworks to further test for ecological validity. Though we believe that our experiments offer important advantages for studying the effects

of UI designs, and our questions in particular, they also suffer from known drawbacks (such as limited generalizability and the relative simplicity of choice environments in the "lab"), which make the use of multiple methods attractive. For this reason, future work will rely on surveys to gather norm-shaping design patterns that are found in the wild and on qualitative coding of design patterns and common "widgets" that are present in social media systems (making use of established patterns) (e.g. *Crumlish and Malone*, 2009) to identify common patterns used by designers.

Though participants were revealing their answers to us (the experimenters), a potential limitation is that the data were obtained from a hypothetical situation. Participants in the rating tasks were asked to evaluate how appropriate they personally felt the photos were to "post"-elicited ratings were for the participants' view of what is (in)appropriate to share on the hypothetical social network site. Though participants did not experience social consequences from revealing information, a user on an social network site would be engaging in the same type of thinking when deciding whether to reveal personal information. Further, studies have suggested that for some decision tasks, participants respond similarly when faced with real and hypothetical consequences (cf. *Wiseman and Levin*, 1996).

2.6 Conclusions

The design of social media interfaces greatly shapes the extent and timing of people's decision to reveal private information. The mechanisms through which interfaces affect disclosure behavior are still poorly understood. Previous research demonstrates how signals of social norms are useful for online social communities to assist newcomers, and can prevent certain behaviors and encourage others *Kraut et al.* (2012). There is also extensive literature on how decision contexts may be altered to nudge behavior towards some target outcome *Acquisti and Grossklags* (2005); *Thaler and Sunstein* (2008).

What we show is that signals embedded in design form social guideposts that shape users' beliefs about what kind, and how much, information is acceptable to reveal and ultimately translate into greater information disclosure. We experimentally unpack the causal pathway by which design patterns affect disclosure: they can modify perceptions of appropriate behavior which, in turn, impact subsequent behavior. We then demonstrate that this shift in perceptions can leave a larger footprint on user behavior than typically anticipated. This is because the shift in perceptions of appropriate behavior also impacts the advice that a user gives others.

Taken together, we identify a powerful cycle of behavior nudging and norm shaping through design patterns. The cycle begins with a design pattern that shapes *perceptions* of what is personally acceptable to do. These perceptions then nudge users to change behavior (in our case, share more about themselves by skipping fewer questions). An increase in sharing of private and information rich data then feeds back into the system and does shape the norms of the community. It does so in two very powerful ways. First, through altered behavior by individual users. Second, it changes the actual norms because those "nudged" users also change what advice they give other newcomers. Taken together, these findings have significant implications for security and privacy.

Social media systems touch the lives of hundreds of millions of end-users daily, and even minor design changes contribute to changes in behavior regarding the sharing of privacy-sensitive information among these users. By failing to account for the norm-setting implications of design choices, or only understanding them through vague intuitions, designers are poorly equipped to model the impact of their choices and the long term consequences to their end-users and sites. End-users and policy makers are similarly blind. In this paper we describe a novel approach to understanding the impact of social media user interfaces on social norms. Our study allows us to test how individuals trade-off the gains from information sharing against norms. As

critically, we are also able to isolate a causal pathway from UI choices designers make to information sharing behavior.

CHAPTER III

Information Wars

3.1 Introduction

Promotion decisions are complex. Managers must anticipate how an individual will perform in a familiar but novel situation. To do so, managers rely on noisy signals of the worker's ability based on his or her current performance. Managers' promotion decisions are dependent on the relative worker performance (*DeVaro*, 2006). However, managers' assessments are not objective and may be influenced by their perceptions of the workers.

Much of the literature on hiring and promotion decisions focuses on how managers' choices may be influenced by an candidate's appearance (*Morrow et al.*, 1990; *Drogosz and Levy*, 1996; *Hosoda et al.*, 2003; *Tews et al.*, 2009), mannerisms (*Bonaccio et al.*, 2016; *Stewart et al.*, 2008), and demographic profile (*Ruderman et al.*, 1996; *London and Stumpf*, 1983). Knowing this, candidates may engage in impression management. In a meta-analysis, *Barrick et al.* (2009) find evidence that how an applicant presents himself or herself influences interviewers' ratings of the applicant. They also find that presentation techniques do not correlate with the actual performance of the applicant. These results suggest that a manager's promotion decision may be susceptible to impression management techniques. However, aside from leveraging these techniques to influence the manager's decision, candidates may also deliberately influence the

manager to change his or her decision through cheap talk.

Cheap talk in this context is the unverifiable and costless communication sent from the candidates to the manager. Consider the case where a candidate knows that the manager is choosing between a coworker and himself or herself for a promotion. The manager has information about the relative performance of these two workers. The candidate may try to influence the manager's beliefs about his or her ability by talking about his or her good work performance. Alternatively, the candidate may choose to talk about his or her coworker's poor work performance. In doing so, the candidate could change the manager's beliefs about his or her current performance and increase the likelihood that he or she may be chosen for the promotion. However, the manager may also discount the importance of these statements since they come from biased individuals. In this paper, I test whether and how cheap talk influences the manager's hiring choice. I further test whether candidates can identify and use the most effective cheap talk.

A manager's decision should not be affected by cheap talk because the statements are biased and unverifiable. However, if cheap talk does influence a manager's decision, there are different channels through which cheap talk can operate. Cheap talk can influence the manager's decision by directly shifting the manager's beliefs of the candidate's relative performance. However, cheap talk can also influence the manager's decision in other ways. A manager often considers traits that are not reflected by the candidate's current job performance (*Reinsch Jr and Gardner*, 2014). An individual's cheap talk may provide the manager with information about such traits. For example, someone who engages in self-promoting behavior may be perceived as more competitive or aggressive (*Rudman*, 1998). The manager's interpretations of the cheap talk may also interact with his or her prior knowledge of the candidate. The manager may interpret a self-promoting statement from a low-performing individual as a lie but may take the same statement at face value if it comes from a

high-performing individual. Thus, cheap talk may influence a manager's decision in ways other than shifting the manager's beliefs about a worker's performance and have different effects given the manager's prior information about the candidate.

I designed a multi-stage experiment to study the impact of cheap talk on managerial decisions. Participants are assigned the role of the "manager" or of the "worker." The managers are tasked with choosing between two workers to hire for an upcoming task similar to the one that both workers completed before. In the first stage, managers receive credible information about the two workers' relative past cheating behavior in an earlier task ("prior performance disclosure"). A worker either cheated "more" than, "less" than, or "as much as" another randomly chosen worker. In the second stage, managers are told the cheap talk responses that the two workers chose in answer to the performance disclosure. The responses may be "sweet" (e.g. the worker reiterates his or her own good qualities), "mean" (e.g. the worker talks about his or her competitor's bad qualities), or "neutral" (e.g. the worker thanks the manager for being considered for the task).

In each stage, managers are incentivized to report their best guess as to the number of times they think one of the two workers cheated in a prior task. Managers report their guess three times: once prior to receiving any information, once after the prior performance disclosure and once more after workers' responses. After these two stages, managers are asked to select either one or neither of the workers to participate in a modified trust game (*Charness and Dufwenberg*, 2006). This setup directly tests the impact of cheap talk on managers' final hiring choices. This design also allows me to test how preceding information about the workers interacts with cheap talk to influence managers.

I find that workers' cheap talk can influence managers' beliefs even though cheap talk offers no objective information. However, the responses only influence beliefs in cases where the managers are told that a worker cheated "less" than or "as much as"

the other worker.

The relative effectiveness of each type of cheap talk depends on the prior performance disclosure. If a worker cheated "as much as" the other worker, the "sweet" response decreases the number of times that the manager believes he or she cheated in the previous task, but the "neutral" and "mean" responses do not. However, all three responses are equally effective in decreasing the number of times that the manager believes the worker cheated when the preceding information is that the worker cheated "less" than the other worker.

Both the prior information disclosures and the subsequent responses have an impact on managers' hiring choices. Managers are more (less) likely to hire the worker when they are told that that worker cheated less (more) than the other worker. The influence of the worker's cheap talk response is dependent on the prior performance disclosure. A worker who cheated "as much as" the other worker is more likely to be hired with a "sweet" response. But a worker who cheated more than the other worker is more likely to be hired with a "mean" response.

Combined, my results suggest different channels through which cheap talk can affect the manager's hiring decisions. A worker's cheap talk can influence the manager's hiring decision by influencing the manager's beliefs about the worker's past cheating behavior. The more times that the manager believes a worker cheated in the past, the less likely he or she is to hire that worker. However, the worker's cheap talk can also directly influence the manager's hiring decision outside of this belief channel. Further, the channel that a cheap talk operates through depends on the type of response and the preceding information. Both the nature of the cheap talk response and the prior performance disclosure affect the manager's beliefs of how many times the worker cheated as well as the subsequent hiring decision.

All three cheap talk responses following the disclosure that the worker cheated "less" than the other worker lower the manager's beliefs about the number of times

the worker cheated. However, a "sweet" response following the "less" performance disclosure has an additional and direct influence on the manager's hiring choice that decreases the likelihood that that worker is hired. When the worker cheated "as much as" the other worker, a "sweet" response lowers the manager's beliefs about the number of times the worker cheated. For this performance disclosure, none of the responses operate outside of beliefs to influence the manager's hiring choice. Lastly, when a worker cheated "more" than the other worker, none of the worker responses change a manager's beliefs about how often the worker cheated. However, the "mean" response still operates outside of the belief channel to increase the likelihood that the worker is hired.

Finally, I also find evidence that some workers are aware that cheap talk's effectiveness depends on the type of preceding information. While most workers choose the "sweet" response regardless of the performance disclosure, a significantly larger fraction of workers choose a "mean" response only when the preceding information is that the worker cheated "more" than his or her competition.

This paper contributes to the understanding of the impact of cheap talk as responses to credible information. Current literature in economics investigates the influence of information on beliefs but has yet to provide a test of how responses to that information may also influence beliefs and choices. The results suggest an interaction between credible information and responses that is novel to the literature. This paper also adds insight into whether everyday individuals can identify and use the most effective rhetoric for a situation. This paper further extends the literature on bargaining games by testing whether individuals can correctly predict the type of response that would work best in different situations.

This paper also makes a direct contribution to the literature on promotions within firms. A manager's promotion decision is often dependent on the manager's perception of the relative effectiveness of the candidates at their current tasks. However,

interactions with candidates can also affect managers' choices by signaling other traits not directly related to current job performance; for example, managers may care that the candidate is trustworthy (*Reinsch Jr and Gardner*, 2014). This paper disentangles and measures the influence that a candidate's persuasive but unverifiable statement has on a manager's decision.

The results of this paper may be extended to other contexts where persuasive communication is used. Although this paper deals primarily with an internal promotion process in which candidates may talk about each other, hiring managers making external hires may also receive similar signals and responses from the applicants. For example, an applicant may choose to talk negatively about his or her current superior. Another potential context is a political election. Candidates may engage in cheap talk in the forms of self-promotion or other-defaming statements when faced with news about themselves or about other candidates. This paper suggests that voters may be susceptible to this form of cheap talk.

More broadly, the results of this paper have implications for both sides of the labor market. While interviews are a valuable tool by which managers may use to help inform their evaluations of candidates, they are also opportunities for candidates to strategically sway managers' decisions. The results of this paper suggest how managers ought to weigh the information acquired during an interview. In particular, managers ought to evaluate how much of what the candidates say is verifiable. For candidates, the results of this paper demonstrate that cheap talk can work, but how well it works depends on what the manager already knows about their past work performance. The paper also suggests that even if candidates are unable to change a manager's impression about their past performance, they may still work to influence the manager's decision in other ways.

3.2 Related literature

A manager's decision-making process is complex. He or she is motivated to hire a candidate who would best fit the new position using incomplete information about candidates' abilities. *Reinsch Jr and Gardner* (2014) model the decision-process by adapting Brunswik's lens model (*Brunswik*, 1955). In this model, a manager can infer a candidate's actual attributes and traits based on cues given off by the candidate, such as his or her work performance. However, these cues are imperfect signals of a candidate's attributes because his or her work performance is a function of the workplace (e.g. coworkers, specific projects, current superior). A manager's interpretation of the candidate's current work performance may also be biased or misinterpreted. For example, a manager's negative first impression of a candidate may cause the manager to discount the positive recommendation from the candidate's current superior. The hiring and promotion decisions literature also finds evidence that decisions are made based on attributes that are not directly relevant to the job (*Hitt and Barr*, 1989; *Ruderman et al.*, 1995; *Drogosz and Levy*, 1996).

The candidate's strategic communication may also influence the hiring and promotion decision processes. *Deshpande et al.* (1994) vary the candidate's level of an attempt to influence a promotion decision by manipulating the candidate's threats to a decision-maker. They find that this threat is only effective when the decision-maker is aware that the candidate has ties in the organization. *Stevens and Kristof* (1995) find evidence of the use of persuasive communication to increase the success of the candidate's interview. In a meta-analysis considering the effects of several influence tactics, *Higgins et al.* (2003) find that two types of persuasive communication, ingratiation and rationality, have a positive effect on the decision-maker's assessment of the individual's performance. Overall, this literature suggests that individuals use persuasive communication in their interaction with their superiors and that some of these tactics are effective. But the literature also suggests that some of these tactics

may be "cheap talk" and that the candidate's true abilities and traits are not correlated with the use of those tactics.[1] Therefore, it is important that managers can identify cheap talk.

The literature of strategic communication in economics has been primarily focused on "cheap talk" in the context of communication between individuals. These messages are "cheap talk" in that they are not enforceable, are costless, and do not directly affect payoffs. However, despite both parties knowing that the cheap talk is nonbinding, it is often effective in increasing coordination and cooperation (*Cooper et al.*, 1992; *Charness*, 2000; *Charness and Grosskopf*, 2004; *Blume and Ortmann*, 2007; *Duffy and Feltovich*, 2002). For example, *Andersson et al.* (2010) vary Proposers' ability to send free-formed messages in an ultimatum game and find higher acceptance rates when communication is available.

Recent studies examine apology as a form of cheap talk.[2] *Abeler et al.* (2010) study cheap talk in the form of costless apologies. Their natural experiment tests the impact that an apology has on customers' neutral or negative evaluation. They find that the cheap talk apology is more effective in causing the customers to remove their evaluations relative to monetary compensations. *Fischbacher and Utikal* (2013) examine the impact that choosing to apologize has on forgiveness. They vary whether the initial harm is intentional or unintentional and give the offender the opportunity to apologize. They find that the victims of the offense are less likely to punish the offender if the offender apologizes after unintentional harm. However, apologies after intended harm result in a worse outcome for the offender than if they say nothing at all.

The negative campaign literature considers how a campaign or a proxy group's (the sponsor's) distribution of negative, targeted information about a rival candidate

[1] See *Barrick et al.* (2009) for a review.

[2] These studies deviate from cheap talk in ultimatum and bargaining game in that there is a history of negative interaction.

affects respondents' evaluations of the target. The information in negative campaigns comes from an organization or an individual motivated to change the decision-makers' beliefs. A sponsor of the campaign broadcasts information to draw attention to his or her opponent's political failures, weaknesses in his or her proposals, or flaws in their characters.[3] In a meta-analysis of 43 studies, *Lau et al.* (2007) find mixed results of the influence of these negative campaigns on actual votes and intent to vote. They find 23 of the 31 relevant studies report that candidates are less liked when they are the target of a negative campaign. However, they also find in 31 of the 40 studies, sponsoring a negative campaign negatively affects the respondents' evaluations of the sponsors. The mixed results suggest that voters (the intended audience of the campaign) both take away information about the target of the campaigns and infer behavioral characteristics about the sponsor (e.g. a candidate "plays dirty"). This is supported by *Dowling and Wichowsky* (2015), *Weber et al.* (2012), and *Brooks and Murov* (2012), who also find that negative ads sponsored by unknown groups, or by groups with unknown connections to the candidate, are more persuasive and produce less backlash for the candidate.

A subset of the negative campaign literature considers the effectiveness of responses to the messages in the negative campaign. *Garramone* (1985) tests the effectiveness of a rebuttal to a negative campaign ad by experimentally manipulating whether participants see a rebuttal commercial after the initial negative ad. She finds that participants have a more negative perception of the benefited candidate if they saw a rebuttal commercial and are less likely to vote for the benefited candidate.

Most relevant is *Craig et al.* (2014), who conduct a two-part experiment where participants read an initial "attack" from a candidate and then read a response from the candidate's opponent. In addition to varying the rhetoric of the "attack" type, they also vary the response types, including "denial," "counterattack," "counterimag-

[3]See *Lau and Rovner* (2009) for a review on the negative campaign literature.

ing," "justification," and "mudslinging." Consistent with some of the prior literature, the attacks lower both the evaluations of the target as well as of the sponsor. Similar to the results in this paper, they also find that responses vary in their effectiveness in changing votes. Interestingly, the combinations of attacks and response do not appear to have a significant impact on the participants' voting choices, but they do have an impact on the overall impression that the participants have of the candidates. Their results suggest that cheap talk operates through different channels depending on the contents of the cheap talk.

3.3 Experimental design

The experiment was conducted over Amazon Mechanical Turk. There was a total of 1,138 participants in the *worker experiment* and 557 participants in the *manager experiment*. No participant participated in both experiments. Table 3.1 displays conditions in the *worker experiment* and the *manager experiment*.

Table 3.1: The two experiments with the associated performance disclosure conditions and response conditions

		Worker experiment			**Manager experiment**		
Response		"Neutral"	"Sweet"	"Mean"	"Neutral"	"Sweet"	"Mean"
Performance disclosure	"A cheated as much"						
	"A cheated less"						
	"A cheated more"						

3.3.1 The worker experiment

This was a between-subject experiment with 3 conditions (performance disclosure: "A cheated as much as" vs. "A cheated less" vs. "A cheated more"). A subject in

this experiment was assigned to either the role of "worker A" or "worker B." This experiment proceeded in two stages. In stage one, the worker completed the *coin-counting task*. In stage two, the worker completed the *response selection task*.

3.3.1.1 Stage one: Coin-counting task

To generate information about the workers' relative work performance, workers first complete a coin-counting task. In this task, the worker evaluated the number of pennies in each of a series of 5 images, regardless of whether the worker was assigned the role of worker A or worker B. There were between 23 to 46 pennies in each image and the presentation order was randomized. The worker received \$0.05 for each correct evaluation for a maximum bonus of \$0.25. After each image, the worker received feedback on the amount that he or she made from the previous answer (e.g. \$0.05 or \$0).

For each of the 5 images, the worker was asked to count the number of pennies in the image, then to report the number on a slider. The worker then "checked" his or her answer by hovering his or her cursor over a coversheet to reveal the answer key. Finally, he or she indicated whether his or her evaluation was correct before going on to the next image. The worker was instructed to check his or her answer only after he or she entered a final answer on the slider.

Since the worker evaluated a total of 5 images, he or she had the opportunity to cheat 0 to 5 times by using an answer key before reporting his or her final count. The worker may do so at no cost to himself or herself and was paid as long as he or she indicated that his or her evaluation was correct.[4]

[4]Other studies have used this type of costless information acquisition in their design. For example, *Dana et al.* (2007) offer dictators the option to costlessly acquire information about payoffs associated with his or her choices.

3.3.1.2 Stage two: Response selection

When making a promotion decision, managers make relative comparisons among candidates (*DeVaro*, 2006). To introduce this relative comparison in the study, in stage two, workers were randomly paired up. I then informed each worker that a participant in another session, the "manager," was deciding to hire the worker, the worker he or she was paired with, or neither of the two workers for a new task (the *one-image coin-counting task*). In this new task, the worker may also cheat by using the answer key and increasing his profit. If the worker chose to cheat by using the answer key, the worker would receive \$0.30 for completing the task as opposed to \$0.21. However, the manager that hire the worker would receive \$0 if the worker chose to cheat (details of this task is available in Section 3.3.3).

The worker then saw the piece of information that would be shared with the manager. The exact information that he or she saw depended on whether he or she was assigned to the role of worker A or worker B and his or her relative performance. The set of possible performance disclosures that could be sent to the manager were:[5]

> "A cheated as much as":
> Worker A (you) cheated as many times as Worker B (the other MTurker) in the earlier HIT.
>
> "A cheated less":
> Worker A (you) cheated less than Worker B (the other MTurker) in the earlier HIT.
>
> "A cheated more":
> Worker A (you) cheated more than Worker B (the other MTurker) in the earlier HIT.

[5]Worker Bs saw that "Worker A (the other MTurker)... Worker B (you)..."

A worker that has been assigned to the role of worker A saw only one of these performance disclosures.[6] For example, if a worker A cheated more times than his or her paired worker B, then worker A saw that the "A cheated more" disclosure would be sent to the manager. Thus, both of the workers and their matched manager shared noisy information about the relative performance of the two workers.

The worker was then given a chance to respond to this performance disclosure. However, neither of the two workers were aware of the choice of the other worker when responding. That is, worker B did not know the choice of worker A. Similarly, worker A did not receive any information about the content of worker B's response.

A worker A chose among one of the following responses to send to the manager:

> "Neutral":
> I will do my best to accurately evaluate the image. Thank you for your consideration!
>
> "Sweet":
> I took my time carefully counting the pennies in each of the images and followed the instructions to the best of my abilities. I will do my best to accurately evaluate the image. Thank you for your consideration!
>
> "Mean":
> Worker B (the other MTurker) did not take his/her time carefully counting the pennies in each of the images and did not follow the instructions to the best of his/her abilities. I will do my best to accurately evaluate the image. Thank you for your consideration!

In contrast, a worker B received one response choice, which was to send the "neutral" response.

[6]Because these sessions were conducted sequentially, to ensure that no deception was involved, worker Bs saw the possible full set of performance disclosures that may be sent to the manager.

This design allowed workers to respond to the information about the relative number of times he or she cheated relative to his or her paired worker.[7] In particular, the response that the worker sent was endogenous to the worker. All of the responses are cheap talk, in that they are costless to send and unverifiable. However, the responses vary in their content. A "sweet" response was unverifiable, but was a statement with information that the worker may have access to (e.g. an individual knows whether or not he or she was careful in the previous task). In contrast, a "mean" response was both unverifiable and the statement contained information that the worker should not know about the other worker.

After selecting a response, the worker completed an exit survey.

3.3.2 The manager experiment

This was a between-subject, experimental design with six conditions (3 (performance disclosure: "A cheated as much as" vs. "A cheated less" vs. "A cheated more") x 3 (response: "neutral" vs. "sweet" vs. "mean")).[8] All subjects in this experiment had the role of "manager."

This experiment proceeded in three stages. In stage one, I described the upcoming tasks. In stage two, I elicited the manager's beliefs of a worker's prior cheating behavior in the *beliefs elicitation task*. Finally, I asked the manager to report his or her hiring preference in the *hiring choice task* in stage three.

3.3.2.1 Stage one: Overview

The manager was randomly assigned a pair of worker A and worker B from the pool of workers in the *worker experiment*. The manager then read about the *coin-counting task* that the workers participated in.

[7] The responses are constructed based on the responses elicited as free responses in a pilot study.

[8] Although the response choice was endogenous in the *worker experiment*, participants in the *manager experiment* were randomly assigned to see a combination of performance disclosure and response conditions.

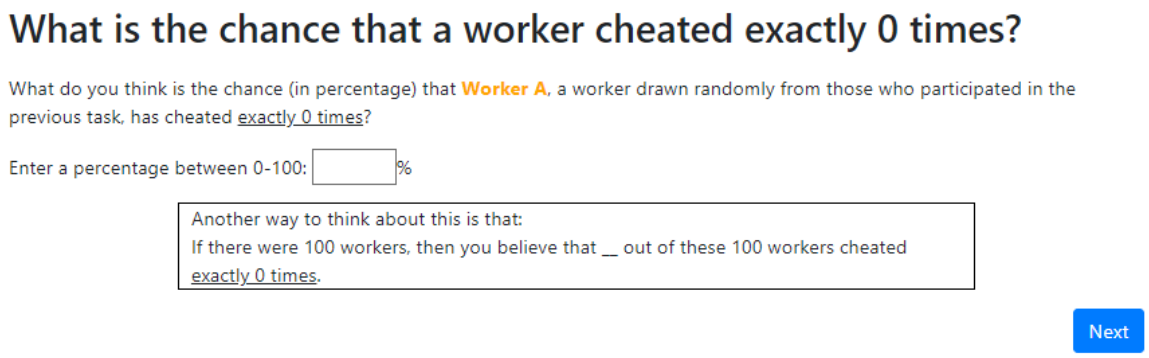

Figure 3.1: An example of what the manager saw in the first round when he or she was asked to report his or her beliefs for each of the number of times that worker A could have cheated, one screen at a time.

The manager's objective in the *beliefs elicitation task* was to report his or her beliefs that worker A cheated exactly 0, 1, 2, 3, 4, and 5 times in the *coin-counting task*. Because prior literature suggests that cues about the candidate's abilities and traits may interact, I elicited the manager's beliefs three times: once before any information, once after seeing the performance disclosure, and once again after seeing the workers' responses.

In round 1, the manager reported his or her prior beliefs that worker A cheated exactly 0, 1, 2, 3, 4, and 5 times in the *coin-counting task*. For round 1 only, I elicited these beliefs in a sequence of screens (see Figure 3.1 for an example of this.)

These beliefs were then used to create an aggregated screen. Beliefs must add up to 100% before the manager could go on to the next screen. The manager could adjust his or her beliefs by typing into the text box. The total sum updated dynamically as managers edited. See Figure 3.2 for an example.

In round 2, I revealed information about the relative performance of worker A and worker B. Specifically, based on the worker A and worker B pair randomly assigned to the manager, he or she saw one of the following:

"A cheated as much as":

Worker A cheated as many times as Worker B in the earlier HIT.

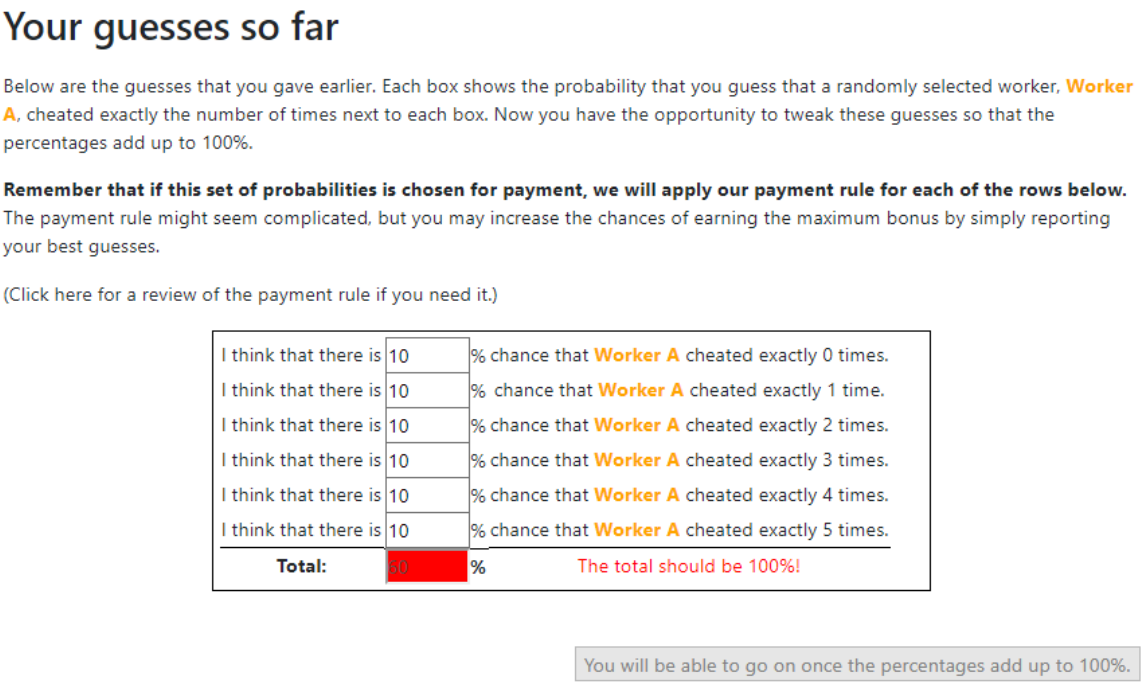

Figure 3.2: An example of the aggregated screen that the manager saw. For each of the possible number of times that worker A cheated, the manager must enter a numerical number. The total must add up to 100% before he or she may go on to the next screen.

"A cheated less":

Worker A cheated less than Worker B in the earlier HIT.

"A cheated more":

Worker A cheated more than Worker B in the earlier HIT.

Afterward, I presented the manager's prior beliefs from round 1 that he or she submitted. The manager then had the chance to revise his or her earlier beliefs. His or her prior beliefs from round 1 and the performance disclosure were also available in a table at the top of the screen.

In round 3, I told the manager that both workers saw and responded to the performance disclosure that the manager saw in round 2. The manager received no additional information about the workers.[9]. The manager saw that worker A sent either the "neutral," "sweet," or "mean" response and worker B sent the "neutral" response.[10]

The manager then had the opportunity to revise the beliefs that he or she reported as the final submission in round 2. The manager's reported beliefs in round 1 were also available in a table at the top of the page.

The manager was incentivized to report his or her beliefs using the crossover mechanism (*Karni*, 2009; *Coutts*, 2019; *Möbius et al.*, 2014). Specifically, following *Berlin and Dargnies* (2016), I stated that:

> Suppose that you think there is a $x\%$ chance that Worker A cheated exactly 0 times. The computer will then generate a random number, y, between 0 and 100.

[9]The manager did not know anything about whether the workers knew what the other worker sent in response to the performance disclosure

[10]These responses were cheap talk in that the contents had no direct payoff implication (*Crawford*, 1998). While the performance disclosure can influence the manager's evaluation of his or her strategy, the responses ought not to influence this. In the current context, the responses were also non-binding and costless for the workers to send. The managers only saw that the workers had sent these responses after seeing the performance disclosure.

1. If y is smaller than (or equal to) x, you will earn a \$0.20 bonus if cheated exactly 0 times.
2. If y is greater than x, you will earn a \$0.20 bonus with $y\%$ chance.

I drew 1 of the 3 rounds of probabilities for payment and implemented this rule for each possible number of times that worker A cheated. Thus, the manager had the opportunity to earn up to a total of \$1.20 from this task. I further told the manager that "[he or she would] get paid most on average when [he or she] honestly reported [his or her] best guesses of the probability for each of the possible number of times Worker A cheated."

The payment rule was available as a hidden element that the manager may click to reveal for each screen the beliefs were elicited. The manager received payment from this stage at the end of the experiment.

3.3.2.2 Stage three: Hiring choice task

In stage three, the manager participated in a modified Trust Game (*Charness and Dufwenberg*, 2006) where he or she may hire either the worker A or worker B that was introduced in stage two. At this point, the manager knew about the relative number of times that worker A and worker B cheated in the previous task (the performance disclosure) and the responses that worker A and worker B sent in response to this disclosure.

I reminded the manager that he or she was hiring one of the two workers for a new task similar to the coin-counting task. I also informed the manager that the hired worker may cheat in this task by using the answer key and receive a larger profit. Whether this hired worker chose to cheat affected the manager's profit. Specifically, the manager's choices and payoffs may be presented in the extensive-form in Figure 3.3.

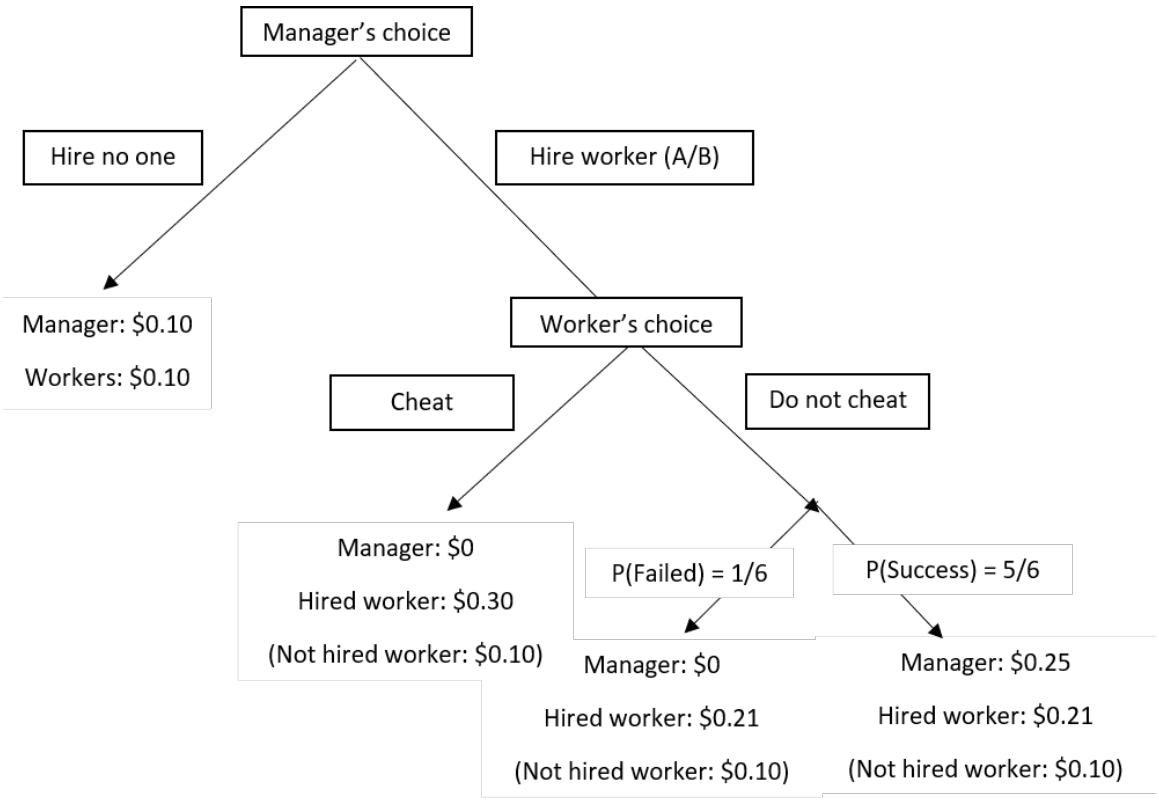

Figure 3.3: The extended form of the modified Trust Game (*Charness and Dufwenberg*, 2006).

If the manager chose not to hire either the worker A or worker B assigned to him or her, then both workers and the manager would receive \$0.10. However, if the manager hired a worker, his or her payoff was then determined by whether the hired worker cheated in this new task by using the answer key.

If the hired worker chose to cheat, then the manager would receive \$0 with 100% certainty and the worker would receive \$0.30. If the hired worker chose not to cheat to complete the task, then the hired worker would receive \$0.21. In this case, the manager's payoffs were \$0 with 1/6 (17%) probability and \$0.25 with 5/6 (83%) probability. The worker who was not hired would receive \$0.10.

The payoffs and the performance disclosures and worker responses from the previous stages were available to the manager as he or she made the decision. The manager also knew that the payment for this stage would be carried out, regardless of whether or not he or she hired a worker, once the sessions for the *one-image coin-counting task* finished.

A feature of this game is that the manager would not be able to infer whether the worker chose to cheat based on the manager's payoff. Further, workers also knew that the manager would not be able to identify their effort. This game mirrored high skilled environments, where the employer may not always directly observe or identify the efforts of their employees.[11]

3.3.3 Final portion of worker experiment (workers' bonus task)

Workers hired by the manager were invited to participate in the *one-image coin-counting task* as described in the *worker experiment* and the *manager experiment*.

[11] If managers and workers were selfish and risk-neutral, then the backward inducted game solution is (*Hire no one*, *Cheat*). However, the manager may take into account his or her beliefs about the number of times that a worker cheated, which may influence the manager's beliefs of the likelihood that the payoffs under "Cheat" and "Do not cheat," were realized (ρ_{cheat}). Specifically, if the manager was risk-neutral and assumed that the worker would behave selfishly, the manager would hire the worker if the manager believed that $\$0.10 < \rho_{cheat} \cdot \$0 + (1 - \rho_{cheat}) \cdot \$0.21$. This is true when $\rho_{cheat} \leq 0.5238$. That is, if the manager believed that there is less than 52.38% chance that the worker would cheat, then the manager would hire the worker.

Specifically, a hired worker was asked to evaluate one image of pennies. He or she completed his or her task when he or she submitted the correct answer. The worker may submit an answer and receive feedback as many times as needed to get to the correct amount. However, the hired worker could also choose to use the answer key. In this case, he or she would receive the answer immediately and would receive a higher payoff.

The hired worker was told the payoffs of his or her choice (as shown in the figure in the previous section) for himself and herself and for the manager that hired them. The payment was paid out to both the hired worker and the manager once the worker completed the task.

3.4 Results

In the following section, I first present the results from the *worker experiment*, then I present the results from the *manager experiment*.

3.4.1 Worker results

All participants evaluate a series of 5 images where they count the number of coins and submit an answer for each of the images. After each submission, they check their answers with an answer key. For each image where they self-report a correct answer, they are paid \$0.05. For the rest of this section, I define cheating as the worker looking at the answer key first before he or she entered an answer or choosing to change his or her answer after looking at the answer key first.

Because all workers participate in this part of the experiment, I do not differentiate between worker As and worker Bs in my analysis.[12] On average, subjects cheated a total of 1.62 times of the 5 possible times. Consistent with the deception literature,

[12]However, there are no differences in the number of times that worker As and worker Bs cheated (t-test, two-tailed p-value = 0.1826).

Figure 3.4: Histogram of the fractions of workers who cheated exactly 0, 1, 2, 3, 4, or 5 times.

a large number of subjects did not cheat, and those that did avoid cheating the maximum amount (see *Rosenbaum et al.* (2014) for a review). 43.50% of participants did not cheat at all, 17.9% cheated 1 time, 8.61% cheated 2 times, 7.03% cheated 3 times, 8.70% cheated 4 times, and 14.24% cheated 5 times. Figure 3.4 presents this visually.

Of the participants who cheated once, 68.27% cheated again in subsequent images. Of the individuals who cheated, 17.93% cheated only on the very last image. Most individuals that cheated began cheating in the very first round (Figure 3.5) and there are no differences in the round that individuals decided to cheat in (Figure 3.6).

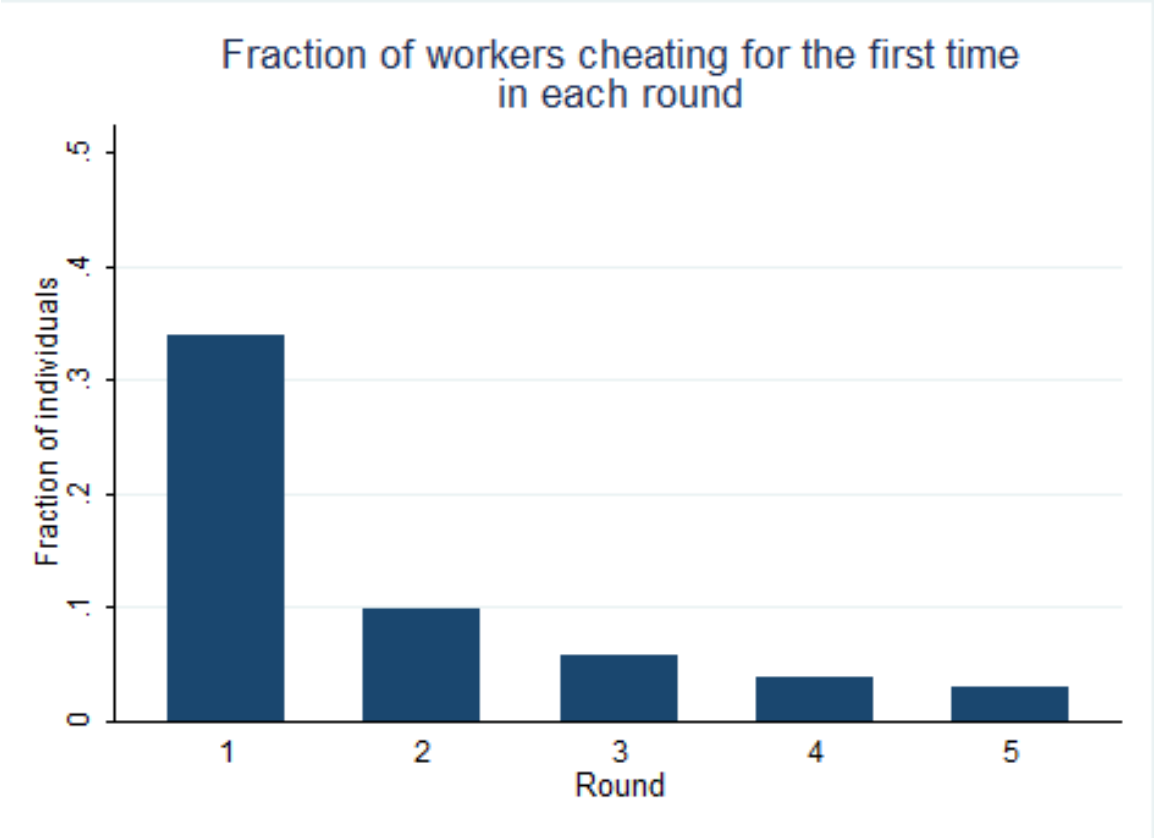

Figure 3.5: Histogram of the fraction of individuals who cheated for the first time in each round.

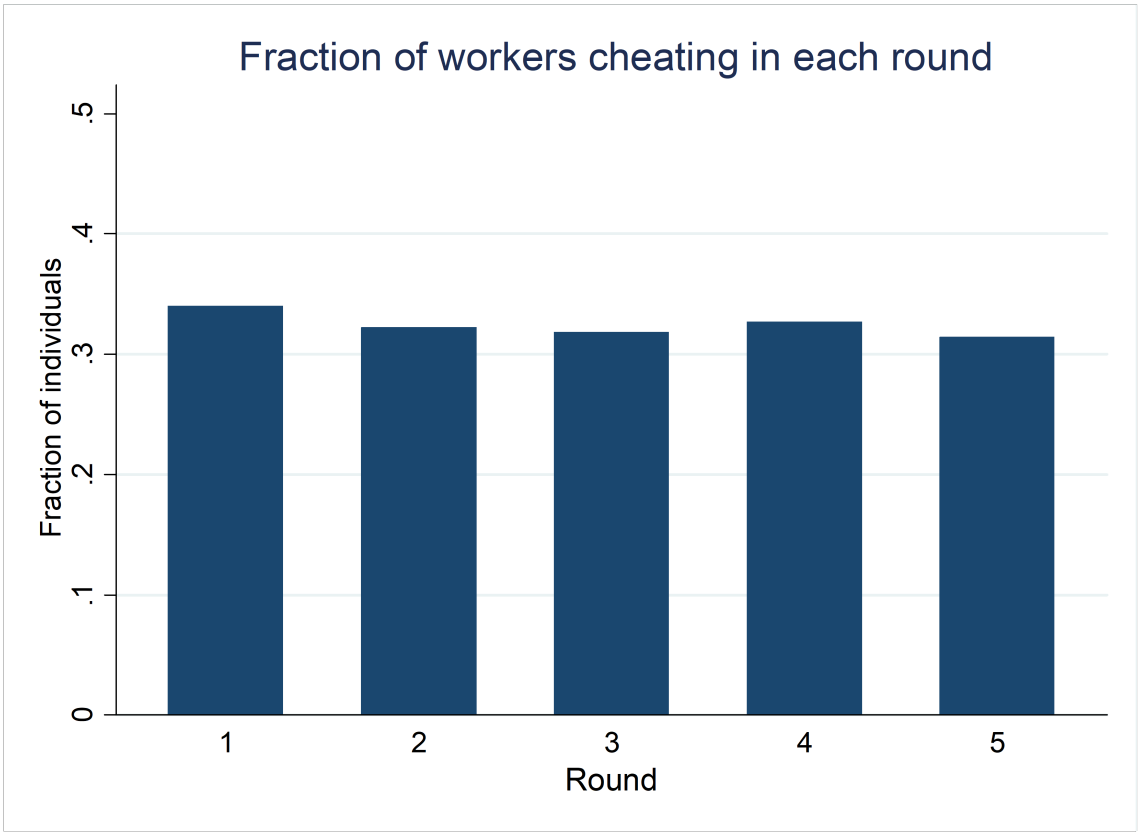

Figure 3.6: Histogram of the fraction of individuals who cheated in each round.

3.4.2 Manager results

I present my results by each stage of the *manager experiment*. I first discuss the results from the *beliefs elicitation task* (stage two), then the results from the *hiring choice task* (stage three). Table 3.2 presents the number of managers in each of the conditions.

Table 3.2: Number of managers in each condition.

	Response conditions		
	Neutral	Sweet	Mean
"A cheated as much as"	36	99	8
"A cheated less"	39	142	11
"A cheated more"	56	142	24

3.4.3 Beliefs elicitation task: Pre-performance disclosure (Round 1)

In round 1, I elicit managers' prior beliefs about the number of times that worker A cheated. Since managers' beliefs in stage 1 are elicited before they receive any information about worker A or worker B, managers' expectations of the number of times A cheated should not differ across the three types of performance disclosure.

As a check, for each manager, I construct his or her expected number of times that worker A cheated by using the priors he or she reported (the manager's prior mean). I then perform an OLS regression of this prior mean on the performance disclosure condition that that manager was in. The OLS regression (Table 3.3) shows no significant differences across the performance disclosure conditions.

3.4.4 Beliefs elicitation task: Post-performance disclosure (Round 2)

In the following section, I report results testing the impact that the performance disclosure had on managers' beliefs.

Table 3.3: OLS regression of managers' reported prior means on performance disclosure conditions.

	Pre-information: Reported prior means
A cheated less	-0.113
	(0.097)
A cheated more	0.003
	(0.099)
Constant	2.452***
	(0.079)
Observations	557
R-squared	0.004

Robust standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1

In stage 2, managers receive information about the number of times that worker A cheated relative to worker B. Before receiving the information, managers' beliefs of the number of times that worker A and worker B cheated ought to be random draws of their beliefs of the worker population's overall cheating behavior. However, in round 2, after receiving information about the relative cheating behavior of worker A and worker B, managers ought to update their beliefs (but not the overall beliefs about the distribution of cheating behavior in the worker population).

If managers update their beliefs based on the content of the information, then the elicited beliefs ought to be different post-performance disclosure. Specifically,

Hypothesis 1 (**Post-information: Update in beliefs**). *Managers' beliefs about the number of times that worker A cheated change after the performance disclosure for "A cheated less" and "A cheated more" but not after the performance disclosure for "A cheated as much as."*

I follow the steps in the pre-performance disclosure analysis to construct the manager's expectation of the number of times that worker A cheated (the posterior mean).

I then construct a variable that is the difference in the prior mean and the posterior mean. When managers learn that "A cheated less," they expect that worker A cheated an average of 0.37 fewer images of the possible 5 images (t-test, p-value < 0.01). When managers learn that "A cheated more," they expect that worker A cheated on an average of 0.35 more images (t-test, p-value < 0.01). However, when managers learn that "A cheated as much as," they also change their expectations of the number of times that A cheated. Managers expect that A cheated on an average of 0.07 more images of the possible 5 images (t-test, p-value < 0.05).[13]

Result 1 (**Post-information: Update in beliefs**). *For the "A cheated less" and the "A cheated more" performance disclosures, the change in the expected number of times cheated pre- and post- performance disclosure is significantly different (p-value* < 0.01*). When managers learn that "A cheated less," they lower their expectations about the number of times that worker A cheated. When managers learn that "A cheated more" than worker B, they increase their expectations about the number of times that worker A cheated (t-test, p-value* < 0.01*). When managers learn that "A cheated as much as," they also increase their expectations about the number of times worker A cheated (p-value* < 0.05*). However, this increase is small in magnitude when compared to the other two performance disclosure conditions.*

3.4.5 Beliefs elicitation task: Post-response (Round 3)

Using the data from the post-response stage of the study, I first report results testing whether managers' beliefs changed post-responses, holding constant each performance disclosure condition. I then test and report whether any of the beliefs changes post-response vary by the type of response and by the performance disclosure preceding the response.

[13]In addition, I compare the absolute difference in the prior means and the posterior means between the "A cheated less" condition and the "A cheated more" condition. I find no asymmetry in the change of expectations across these two conditions (t-test, p-value = 0.4226). Additional tests of asymmetry in beliefs update are available in the Appendix.

3.4.5.1 Changes in managers' post-response beliefs collapsed across all response conditions

After receiving the performance disclosure, managers see the workers' responses. Given that these responses are "cheap talk" in the informational sense, I expect that

Hypothesis 2 (**Post-response: Update in beliefs**). *Managers do not update beliefs after receiving workers' responses.*

I again construct a variable that is the difference of the posterior means before the manager sees the responses and after he or she sees the responses. I perform t-tests for the difference for each of the performance disclosure conditions. The t-tests are significant for the disclosure conditions "A cheated as much as" (p-value < 0.01) and "A cheated less" (p-value < 0.01), but not for the disclosure condition "A cheated more" (p-value $= 0.7099$).

Result 2 (**Post-response: Update in beliefs**). *Managers revise their expectations after receiving the responses if the responses are preceded by either the "A cheated as much as" disclosure or the "A cheated less" disclosure. However, managers do not revise their expectations after receiving the responses the preceding disclosure is that "A cheated more."*

3.4.5.2 Changes in managers' post-response beliefs by response conditions

In the previous section, I find that for at least two of the performance disclosure conditions, managers revised their expectations about the number of times that worker A cheated. In this current section, I test the impact that the responses have in changing prior beliefs, holding constant the preceding performance disclosure.

Figure 3.7 presents the histogram of the changes in managers' beliefs, by the preceding performance disclosure and the subsequent response. For example, the

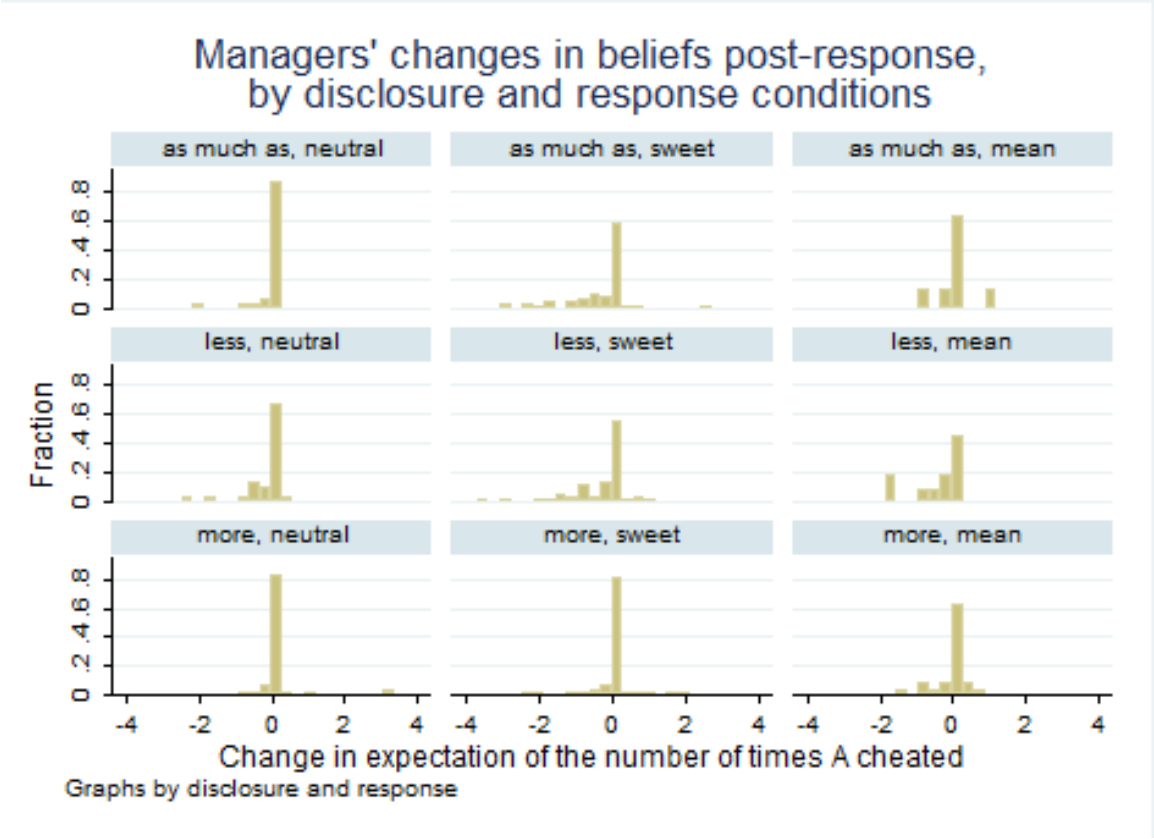

Figure 3.7: Changes in managers' beliefs of the number of times that worker A cheated post-response by the performance disclosure condition and by the response condition.

histograms in the first row represent the changes in managers' expectations about the number of times that worker A cheated when the preceding performance disclosure is "A cheated as much as." The single histogram in the first column of the first row presents the change in managers' expectations for when worker A's response is a "neutral" response.

If the specific type of response did not affect managers' expectations of the number of times that worker A cheated, then there should be no differences across the three histograms within each row. However, visually, it appears that the type of responses did have an impact on managers' expectations of the number of times that worker A cheated.

Table 3.4 reports the OLS regressions of managers' posterior means on dummy variables for the "sweet" response condition and for the "mean" response condition. Column 1 presents the regression for managers in the "A cheated as much as" performance disclosure condition. When the preceding information is "A cheated as much as," testing managers' changes in expectations of the number of times that worker A cheated against no change in expectations, only the "sweet" response significantly lowers managers' expectations (Wald test, p-value < 0.01). The "sweet" response also lowers expectations of the number of times A cheated more than the "neutral" response (Wald test, p-value < 0.01) as well as relative to the "mean" response (Wald test, p-value < 0.05).

Column 2 presents the regression for managers in the "A cheated less" performance disclosure condition. All three responses significantly lower managers' expectations of the number of times that worker A cheated relative to no change (Wald tests, p-value < 0.05 for the "neutral" response; p-value < 0.01 for the "sweet" response; p-value < 0.05 the "mean" response). However, neither the "sweet" nor the "mean" responses lead to greater changes in managers' expectations relative to the "neutral" response.

When the preceding information is that "A cheated more" (Column 3), none of the responses are effective in changing managers' expectations of the number of times that worker A cheated. Overall, I find that

Result 2a (**Post-response: Update in beliefs**). *Whether and by how much the response changes beliefs about the number of times that worker A cheated vary by the type of response and by which performance disclosure precedes the response.*

Table 3.4: OLS regressions of the change in the reported posterior means on the dummy variable each of the response conditions, holding constant the performance disclosure condition within each column.

	Post-performance disclosure: A cheated as much as	Post-performance disclosure: A cheated less	Post-performance disclosure: A cheated more
Sweet	-0.269***	-0.117	-0.127
	(0.101)	(0.098)	(0.092)
Mean	0.129	-0.248	-0.188
	(0.173)	(0.200)	(0.121)
Constant	-0.090	-0.188**	0.113
	(0.063)	(0.079)	(0.087)
Observations	143	192	222
R-squared	0.039	0.008	0.018
Wald tests: P-values of F-statistic			
Sweet vs. Mean	0.0286	0.4968	0.4943
Neutral = 0	0.1521	0.0186	0.1932
Sweet = 0	0.0000	0.0000	0.6700
Mean = 0	0.8109	0.0184	0.3710

Robust standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1

3.4.5.3 Managers' final beliefs by information conditions and response conditions

Finally, to examine the overall impact of the performance disclosure and responses on managers' final expectations, I restrict the data to managers' beliefs reported in the last round (round 3). I construct managers' expectations of the number of times that worker A cheated using the same method as in the previous sections. Figure 3.8 displays the posterior means in the final round by the performance disclosure condition (rows) and by the response condition (columns).

Visually, managers' expectations of the number of times that worker A cheated appear to vary by both the performance disclosure condition and the response condition. For example, a greater fraction of managers expected that worker A cheated fewer times in the "A cheated less" condition relative to the other two information conditions. Further, the distributions of managers' expectations appear to also differ across the response conditions.

To test the overall impact that the performance disclosure conditions and the response conditions have on managers' expectations, I perform an OLS regression of the posterior means on the dummy variables for the response conditions, the dummy variables for the performance disclosure conditions, and their interaction terms. The

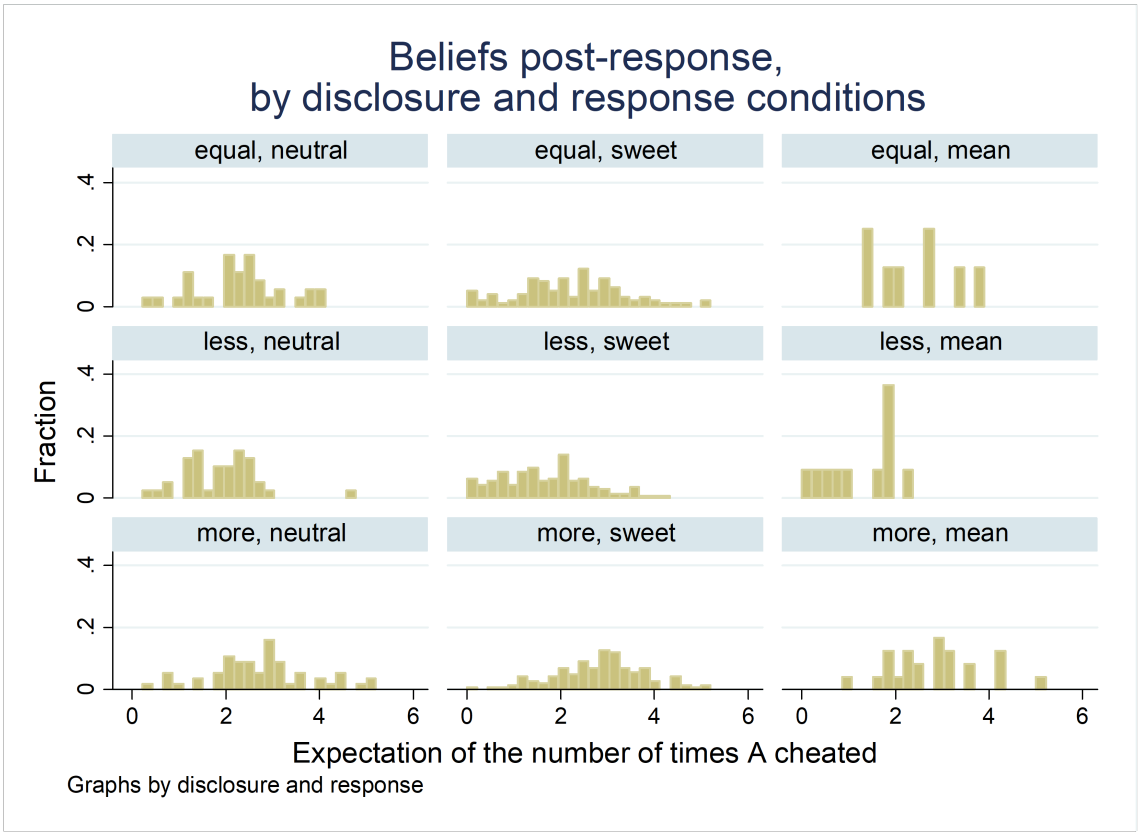

Figure 3.8: Managers' final expectations of the number of times worker A cheated by the performance disclosure condition and the response condition.

result of this OLS regression is available in the Appendix (Table B.2).

Holding constant the performance disclosure condition, I test to see if managers' final expectations of the number of times that worker A cheated differ across response conditions. For the "A cheated as much as" performance disclosure condition the Wald tests testing for these differences across responses are insignificant (p-values for "neutral" versus "sweet" is 0.7143, for "neutral" versus "mean" is 0.7206, and for "sweet" versus "mean" is 0.5505).

The results are different for the "A cheated less" performance disclosure. While the impact of "neutral" versus "sweet" are not significantly different (Wald test, p-value = 0.1181), there are slight differences in the impact of "neutral" versus "mean" (Wald test, p-value = 0.0141) and "sweet" versus "mean" (Wald test, p-value = 0.0983).

Lastly, none of the responses appear to have an influence in the "A cheated more" performance disclosure condition. The Wald tests testing for these differences are insignificant (p-values for "neutral" versus "sweet" is 0.4343, for "neutral" versus "mean" is 0.6450, and for "sweet" versus "mean" is 0.9138).

Overall, the effectiveness of the different types of responses at changing beliefs varies by the preceding performance disclosure.

3.4.6 Managers' hiring choices

After eliciting their beliefs of the number of times that workers A cheated, managers make a hiring choice. Using the data from this hiring choice, I first report results testing whether the performance disclosure conditions have an impact on managers' hiring choices. I then test whether any of the responses increase the likelihood that a manager hires worker A. Finally, I test the impact that the performance disclosure, response, and managers' beliefs have on managers' hiring choices.

3.4.6.1 Managers' hiring choices collapsed across all response conditions

Managers have the choice of hiring either worker A, worker B, or neither of the two workers to participate in the *workers' bonus task*. I create a dummy variable to indicate whether a manager chooses to hire worker A. Given that managers lower their expectations of the number of times that worker A cheated the most when the performance disclosure is that "A cheated less" than the other worker, I expect that

Hypothesis 3 (**Managers' hiring choices: Impact of performance disclosure**). *Managers are most likely to hire worker A if the performance disclosure is that "A cheated less."*

Table 3.5 presents a logistic regression of the dummy variable for hiring worker A on dummy variables for the performance disclosure conditions. The dummy variables on the performance disclosure are significant. In particular, the performance disclosure "A cheated less" increases the likelihood that the manager hires worker A (p-value < 0.01) relative to the disclosure that "A cheated as much as." In contrast, "A cheated more" decreases the likelihood that the manager hires worker A (p-value < 0.01) relative to the disclosure that "A cheated as much as." Further, "A cheated less" also significantly increases the likelihood that worker A is hired, relative to "A cheated more" (Wald test, p-value < 0.01

Result 3 (**Managers' hiring choices: Impact of performance disclosure**). *When the performance disclosure is "A cheated less," managers are more likely to hire A relative to either of the other performance disclosure conditions. When the performance disclosure is "A cheated more," managers are less likely to hire worker A relative to either of the other performance disclosure conditions.*

).

Table 3.5: Logistic regression of the dummy variable for hiring A on the dummy variables of the performance disclosure conditions.

	Dummy variable for hiring A
A cheated less	0.888***
	(0.240)
A cheated more	-1.873***
	(0.246)
Constant	0.296*
	(0.169)
Observations	557
Wald test: P-value	
A cheated less vs. A cheated more	0.0000

Robust standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1

3.4.6.2 Managers' hiring choices by the response conditions

In the previous section, I test the impact that the performance disclosure has on the likelihood that worker A is hired, collapsed across all of the response conditions. In the current section, I hold constant the preceding performance disclosure condition and test whether the different response conditions have differing effects on the likelihood that worker A is hired.

I hypothesize that

Hypothesis 4 (**Managers' hiring choices: Impact of response**). *Managers' likelihoods of hiring A differ across response conditions, given the performance disclosure.*

Figure 3.9 presents the hiring choices of managers for each combination of performance disclosure condition and response condition. The y-axis indicates the fractions of managers who hire neither of the workers ("None"), to hire worker A ("A"), or to hire worker B ("B"). Histograms in the same row are choices of managers in the same

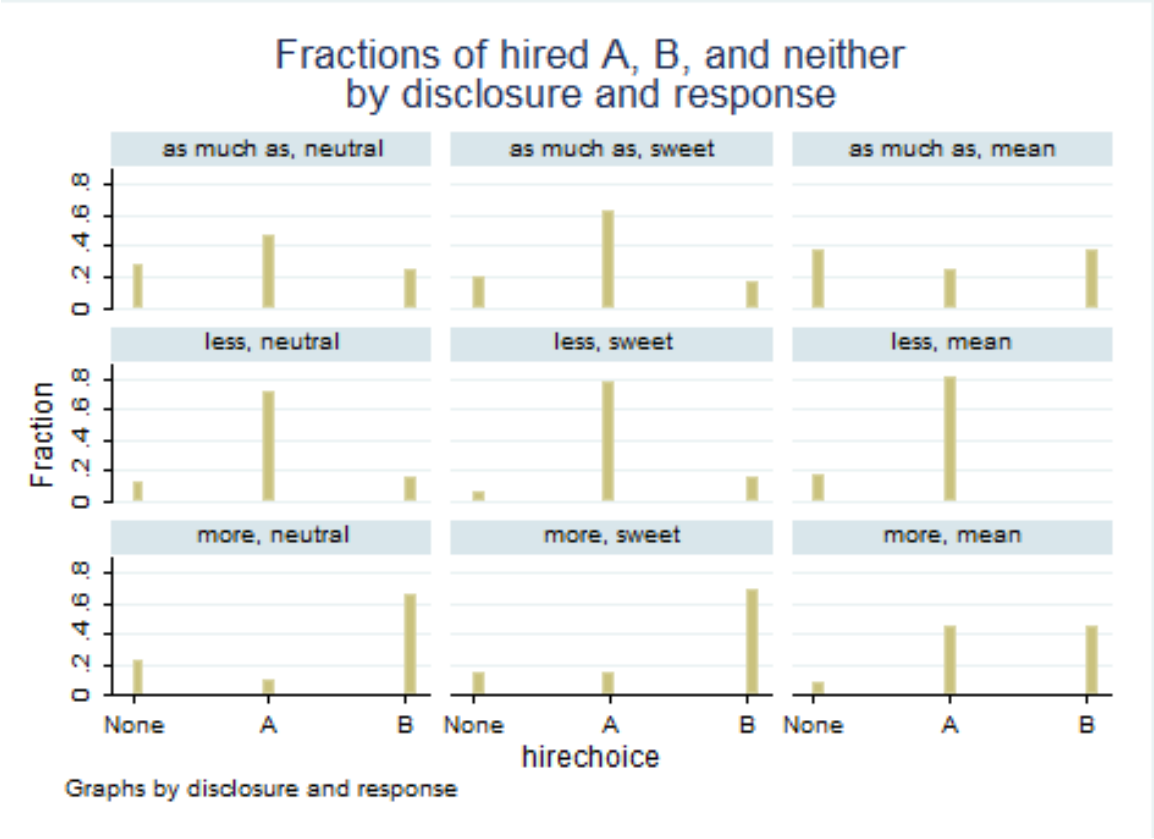

Figure 3.9: Managers' hiring choices for each combination of performance disclosure condition and response condition.

performance disclosure conditions. For example, the first row is the choices of managers in the performance disclosure condition "A cheated as much as." Histograms in the same column are choices of managers in the same response conditions. Thus, the upper-left most histogram shows the choices of managers who are in both the "A cheated as much as" performance disclosure condition and the "neutral" response condition.

Visually, it appears that the different response conditions have different impacts on the likelihood that a manager hires worker A. If the responses have no impact on managers' hiring choices, then the fraction of hiring worker A should not differ across the three columns for each of the rows. However, for both the "A cheated as much as" disclosure condition (Row 1) and "A cheated more" disclosure condition (Row 2),

it appears that the fractions of managers choosing to hire A differ across the response conditions. Using the same hiring A dummy as before, I perform a series of logistics regressions of the hiring A dummy on dummy variables for the response conditions. I report the results of these regressions in Table 3.6.

Looking first at the hiring choices in the "A cheated as much as" disclosure condition (Column 1), I find that the "sweet" response marginally increases the likelihood that A is hired relative to the "neutral" response (p-value < 0.10) and that the "sweet" response increases the likelihood that A is hired relative to the "mean" response as well (p-value < 0.05). When the preceding information is that "A cheated less," none of the response conditions are more effective than the other in increasing the likelihood of hiring worker A (Column 2). Finally, when the preceding information is that "A cheated more" (Column 3), the "mean" response increases the likelihood that worker A is hired over the "neutral" response (p-value < 0.01) as well as the "sweet" response (p-value < 0.01).

Table 3.6: Logistic regressions of the dummy variable for hiring A on the dummy variables for the response condition.

	(1) A cheated as much as: Dummy variable for hiring A	(2) A cheated less: Dummy variable for hiring A	(3) A cheated more: Dummy variable for hiring A
Sweet	0.671*	0.300	0.369
	(0.395)	(0.410)	(0.494)
Mean	-0.987	0.570	1.953***
	(0.885)	(0.861)	(0.597)
Constant	-0.111	0.934***	-2.120***
	(0.335)	(0.357)	(0.433)
Observations	143	192	222
Wald tests: P-values of F-statistic			
Sweet vs. Mean	0.0499	0.7393	0.001

Robust standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1

Overall, these results suggest that

Result 4 (Managers' hiring choices: Impact of response). *Managers' likeli-*

hoods of hiring worker A differ across response conditions. However, how responses differ in impact depends on the preceding performance disclosure. The most effective response when the performance disclosure is that "A cheated as much as" is a "sweet" response. When the performance disclosure is that "A cheated more" the most effective response is a "mean" response. There is nothing that A can say that will either help or hurt him or her if the preceding information is that "A cheated less."

Because managers can choose to hire neither of the workers or the other worker (B), it could be that the responses affect not only the likelihood that worker A is hired, but also influence the likelihood of managers hiring worker B. To test this, I performed a series of multinomial logistic regression for each of the performance disclosure conditions. Table 3.7 presents the results of these regressions.

Looking only at the "A cheated as much as" performance disclosure condition, the relative odds of hiring worker A versus hiring neither of the workers do not differ significantly across the responses (Column 1). The relative log odds of hiring worker B versus hiring neither of the workers also do not differ significantly across the responses as well (Column 2).

When the preceding performance disclosure is that "A cheated less," the relative log odds of hiring worker A versus hiring neither of the workers do not differ significantly across the responses (Column 3). However, for the same information condition, the relative log odds of hiring B versus hiring neither of the workers decrease by -14.5887 if response changes from the "neutral" response to the "mean" response (p-value < 0.01). In addition, the effects of the "sweet" and the "mean" response are significantly different from each other in predicting the likelihood of hiring worker B (p-value < 0.01). In other words, although worker A's response does not increase the likelihood that he or she would be hired in this condition, he or she can decrease the likelihood of the manager hiring worker B with a "mean" response.

Finally, when the performance disclosure condition is that "A cheated more,"

the relative odds of hiring A versus hiring neither of the workers increase by 2.4779 if response changes from the "neutral" response to the "mean" response (p-value < 0.01). That is, worker A's response can influence the likelihood of the manager hiring worker A even when the performance disclosure is something "bad" about himself or herself. In addition, the effects of the "sweet" response and the "mean" response are different from each other in predicting the likelihood of hiring worker A (p-value < 0.05). However, the relative log odds of hiring worker B over hiring neither of the workers do not differ significantly across the responses.

Overall, it appears that both performance disclosure and responses have an impact on changing managers' beliefs as well as on managers' hiring choices. In particular, the impacts of responses on changing managers' beliefs and on influencing managers' hiring choices are dependent on the preceding performance disclosure. For example, a "sweet" response only works best to change beliefs if it is preceded by the performance disclosure that "A cheated as much as."[14]

3.4.7 Impact of responses on managers' hiring choices outside of beliefs

The results from the previous two sections suggest that different types of responses may interact with the prior information that managers have to work differently to influence managers' hiring choices. Section 3.4.5.3 shows that certain responses are effective in changing managers' beliefs when preceded by either the "A cheated as much as" or "A cheated less than." If managers' hiring choices depend only on beliefs, then the influence of the responses ought to be entirely through the beliefs.

[14]With the data from the *workers' bonus task*, I also test whether managers' are correct in their choices. Figure B.1 presents the fractions of hired workers who either cheated or not in the *workers' bonus task*. Visually, there appear to be some differences in the fractions of workers who cheated across the different responses. For performance disclosure conditions "A cheated as much as" and "A cheated more," more workers cheated in the bonus task when the worker's response is "mean" relative to the other two responses. I perform a series of logistic regressions on the likelihood that the hired worker cheats on the task based on the responses and disclosure condition he or she is in (Table B.1. I find that while the coefficients are consistent with the visualization, there are no statistically significant differences in the likelihoods that the hired worker cheats, given that particular worker's response. However, this may be due to the small fraction of workers in some of these conditions.

Table 3.7: Multinomial logistic regressions of the hiring choices on the dummy variables for response conditions, for each of the performance disclosure conditions.

	(1)	(2)	(3)	(4)	(5)	(6)
	A cheated as much as		A cheated less		A cheated more	
Outcome:	Hiring A	Hiring B	Hiring A	Hiring B	Hiring A	Hiring B
Sweet	0.617	-0.118	0.675	0.606	0.727	0.458
	(0.476)	(0.571)	(0.589)	(0.717)	(0.582)	(0.400)
Mean	-0.936	0.105	-0.219	-14.589***	2.478***	0.659
	(1.000)	(0.940)	(0.922)	(0.933)	(0.916)	(0.835)
Constant	0.531	-0.105	1.723***	0.182	-0.773	1.046***
	(0.400)	(0.461)	(0.487)	(0.607)	(0.495)	(0.323)
Observations	143	143	192	192	222	222
Wald tests: P-values of F-statistic						
Sweet vs. Mean	0.1027	0.8011	0.2932	0.0000	0.0346	0.8033

Robust standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1

However, when managers are making hiring and promotion decisions, they also place weight on traits of the candidates that are non-ability related (e.g. trustworthiness). In the current context, managers may infer these characteristics based on the type of response that worker A chooses to send. For example, a worker who chooses to send a "sweet" response may be one that frequently engages in self-promotion, which can be perceived as a negative characteristic depending on what the manager knows about the worker (*Rudman*, 1998). In contrast, a worker who chooses to send a "mean" response might be perceived as someone with poor interpersonal skills, which is also a trait important in a promotion decision (*Ruderman and Ohlott*, 1994). Therefore, it is possible that managers' hiring choices can also depend on what they infer from the worker's choice of response in addition to the responses' direct impact on beliefs.

To examine the impact that response and beliefs have on managers' hiring choice, I perform a series of logistic regressions, but restrict my analysis to managers whose

beliefs did not change after seeing workers' responses. In other words, for these managers, the workers' responses do not have an effect on their beliefs. This means that if managers' hiring choices depend entirely on managers' beliefs of the worker's past performance, then the likelihood that managers hire worker A should not differ across the responses. Conversely, if the likelihood of managers hiring worker A varies across the responses for the given performance disclosure, then there must be another channel through which responses affect managers' hiring decisions.

Table 3.8: Logistic regressions of the hiring choices of managers whose beliefs did not change on the dummy variables for response conditions, for each of the performance disclosure conditions.

	(1) A cheated as much as: Dummy variable for hiring A	(2) A cheated less: Dummy variable for hiring A	(3) A cheated more: Dummy variable for hiring A
Sweet	0.522	-1.113*	0.067
	(0.478)	(0.669)	(0.563)
Mean	-0.875	-0.894	1.531**
	(1.225)	(1.315)	(0.745)
Constant	-0.223	1.992***	-2.001***
	(0.390)	(0.618)	(0.478)
Observations	85	104	158
	Wald tests: P-values of F-statistic		
Sweet vs. Mean	0.242	0.854	0.023

Robust standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1

Table 3.8 presents the logistic regressions of the dummy variable of hiring worker A on the dummy variables for the response conditions for each of the performance disclosure conditions. These regressions include only managers whose beliefs are not affected by the response conditions (62.30% of managers). When the preceding information is that worker "A cheated as much as" the other worker (Column 1), there are no significant coefficients on the response conditions and the Wald test testing for differences between the "sweet" and the "mean" response is also not significant (Wald test, p-value = 0.242). That is, any differences in managers' hiring choices within this performance disclosure condition appear to be attributed to differences in

managers' beliefs.

When the preceding information is that worker "A cheated less" than the other worker (Column 2), the sweet response ("Sweet") marginally decreases the likelihood that worker A is hired relative to a neutral response (p-value < 0.10). However, there are no differences in the impact of the "sweet" or the "mean" response on the likelihood that managers hire worker A (Wald test, p-value $= 0.854$). In other words, when the preceding performance disclosure is that worker "A cheated less" than the other worker, managers' decisions are affected by the "sweet" response in a way that is not captured by their beliefs about worker A's past cheating behavior.

Column 3 presents the results for when the preceding information is that worker "A cheated more" than the other worker. The "mean" response increases the likelihood that worker A is hired relative to the "neutral" response (p-value < 0.05). Further, the differences in impact on hiring choices are also significant for the "sweet" response and the "mean" response (Wald test, p-value $= 0.023$). This suggests that the "mean" response significantly influences managers' decisions outside of changing their beliefs.

Altogether, the results in this section is evidence that responses operate outside of the beliefs channel to directly influence managers' decisions. However, as with their impact on managers' beliefs, the impact that responses have on managers' hiring choices outside of changing their beliefs depends on what managers know about worker A prior to receiving the workers' responses.

3.4.7.1 Workers' response choice

Recall that worker As are making a choice about the type of responses they want to send their managers. If the workers are able to identify and use the response that would best change hiring outcomes in their favor, then workers ought to

1. Choose "sweet" response when the preceding information is that "A cheated as much as"

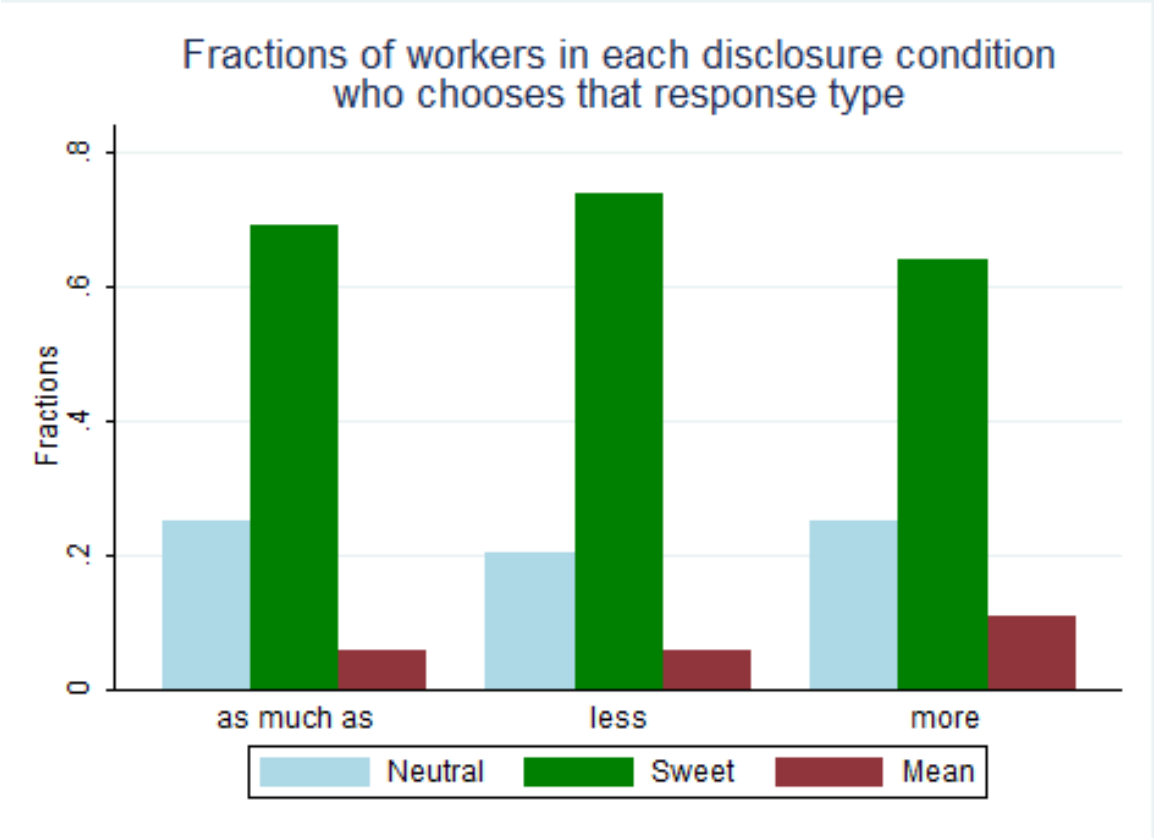

Figure 3.10: Bar graph of the fractions of workers in each performance disclosure condition, choosing that particular response.

2. Choose "mean" response when the preceding information is that "A cheated more"

Figure 3.10 displays the fractions of workers in each performance disclosure condition choosing each type of response.[15] The bars are grouped by the performance disclosure condition. The blue bars represent the workers whose response is the "neutral" response, the green bars represent the workers whose response is the "sweet" response, and the red bars represent the workers whose response is the "mean" response.

The "sweet" response is the most frequently selected response across all three performance disclosure conditions (test of proportions, two-tailed, p-value < 0.01).

[15] A total of 557 randomly paired workers are used in the results for this section.

However, visually, it appears that there are some differences in the fractions of workers who choose the "neutral" and "mean" response across the performance disclosure conditions.[16]

To test whether the fractions of workers choosing "sweet" and choosing "mean" responses differ across performance disclosure conditions, I perform a series of tests of proportions. I find that there is a higher proportion of workers choosing the "sweet" response in the "A cheated less" condition relative to the "A cheated more" condition (test of proportions, two-tailed, p-value < 0.05), but no significant differences when compared with the "A cheated as much as" condition. In addition, the proportion of workers choosing the "mean" response is higher than the proportion of workers choosing the same response in both the "A cheated less" condition and the "A cheated as much as" condition (test of proportions, two-tailed, p-value < 0.10).

Overall, workers choose the "sweet" response most frequently, regardless of the prior performance disclosure. While this is the effective response when the preceding performance disclosure is "A cheated as much as" and "A cheated less," the most effective response when the preceding performance disclosure is "A cheated more" is the "mean" response. There is some evidence that a fraction of the workers are aware that the most effective choice is the "mean" response. This is seen in the smaller proportion of workers choosing the "sweet" response and the larger proportion of workers choosing the "mean" response only in the "A cheated more" disclosure condition.

3.5 Conclusion

This paper tests the impact of cheap talk on an individual's beliefs and choices in hiring. I find that managers update their beliefs in reaction to workers' chosen cheap

[16] All workers also report their belief of what would be the most effective message to send, and this is consistent with the results in this section. See Appendix for details of the analysis.

talk responses. The degree of impact that these responses have on the managers' beliefs is dependent on prior knowledge of a candidate's past performance.

In cases where a manager has been told that a worker cheated "as much as" another worker, "sweet" responses can significantly lower the manager's beliefs about how many times that worker cheated. While no specific response outperforms the other, all three response types also significantly lower the number of images that the manager believes worker A cheated on if the preceding information disclosure is that worker A cheated "less" than the other worker. Lastly, worker A is most likely to be hired if he or she chooses a "mean" response if the preceding information is that he or she cheated "more" than the other worker.

Further, this paper also finds evidence that there are different channels through which the responses operate. A response may change the managers' beliefs of the number of times that the worker cheated in a previous task. However, managers' are also influenced by the worker's choice of response in a way that is outside of changing managers' beliefs about the worker's past cheating behavior. This is because managers infer behavioral qualities about the candidates based on the type of response that a worker sends. For example, a manager may perceive a worker who gives a "mean" response as either "petty" or "direct" depending on the manager's prior knowledge about that worker. These results in this paper suggest that candidates may leverage cheap talk to influence managers' promotion decisions.

This paper also tests the ability of an everyday individual in identifying and using effective rhetoric to persuade another. The "sweet" response is the most effective response in changing managers' beliefs when the prior information is in the worker's favor ("A cheated less") or not informative ("A cheated as much as"). However, when the preceding information is unfavorable for the worker (e.g. "A cheated more"), the "mean" response best increases the chance that the worker is hired. Workers most commonly chose the "sweet" response across all three information disclosure

conditions, however, there is also a larger number of workers choosing the "mean" response in the "A cheated more" disclosure condition relative to the other two.

Recent work in economics has begun incorporating non-pecuniary considerations when modeling how people interpret and react to communication with other individuals. This paper adds to the literature by demonstrating that the impact that cheap talk has on decision-makers can occur through a beliefs channel as well as an alternative channel. In particular, decision-makers may infer information about what "type" of a person an individual is if that individual chooses to engage in self-promotional communication or to engage in other-defaming communication- an evaluation not based no verifiable and directly relevant information. This paper shows that relative past performance disclosure, a noisy but verifiable piece of information, is a determinant for the effects that cheap talk will have on a decision-makers' choice.

This study has practical implications for managers and other decision-makers. Managers often rely on a worker's relative past performances when making a promotion decision, but these performances are often noisy signals of the candidate's actual ability and effort. These results suggest that non-verifiable communication from candidates (e.g. during interviews) can influence managers. Further, for candidates, these results suggest cheap talk as an additional persuasive technique that they may leverage during interactions with the decision-maker. However, a candidate must also take into account the information that a manager have about his or her relative performance; depending on the manager's prior knowledge, the different types of cheap talk may have different effects.

These types of persuasive, interpersonal communication are frequently used in broader contexts. Public figures, political parties, and companies regularly utilize social media platforms such as Twitter, Facebook, and Instagram to communicate directly with their audience, and they often leverage cheap talk to influence audience choices. A company that has been nominated for a prestigious award often capitalize

on this disclosure by using a "sweet" statement on their social media. Conversely, a company whose competitor is involved in a scandal may best respond to this disclosure with a "mean" statement. Cheap talk may influence the audience's beliefs and cause the audience to disregard or to under-weigh the prior credible disclosure. The results of this paper demonstrate the susceptibility of everyday individuals to cheap talk and illustrate how these cheap talk choices may sway public opinions.

APPENDICES

APPENDIX A

Rhetoric Matters: A Social Norms Explanation for the Anomaly of Framing

A.1 Additional Tables

Table A.1: Elicited norms for tax-framed and neutrally-framed Democrats

Final allocation ("Keep")	Tax framed Democratic norm ratings								Neutrally framed Democratic norm ratings							
Panel A: Endowment - 0 Tokens																
	Action	**Mean**	- - -	- -	-	+	+ +	+ + +	**Action**	**Mean**	- - -	- -	-	+	+ +	+ + +
0, 10	"Do not involve government"	**-0.05**	29.55%	12.88%	11.36%	7.58%	13.64%	25.00%	"Do nothing"	**0.54**	7.89%	7.02%	4.39%	8.77%	16.67%	55.26%
1, 9	"Take tax transfer 1"	**0.08**	18.94%	14.39%	11.36%	12.88%	17.42%	25.00%	"Take 1"	**0.35**	6.14%	9.65%	15.79%	14.91%	16.67%	36.84%
2, 8	"Take tax transfer 2"	**0.17**	6.82%	15.91%	16.67%	20.45%	18.18%	21.97%	"Take 2"	**0.33**	5.26%	9.65%	18.42%	13.16%	20.18%	33.33%
3, 7	"Take tax transfer 3"	**0.25**	1.52%	9.85%	20.45%	25.76%	28.79%	13.64%	"Take 3"	**0.29**	6.14%	11.40%	14.91%	14.91%	27.19%	25.44%
4, 6	"Take tax transfer 4"	**0.29**	0.76%	5.30%	19.70%	33.33%	27.27%	13.64%	"Take 4"	**0.29**	5.26%	11.40%	10.53%	19.30%	35.96%	17.54%
5, 5	"Take tax transfer 5"	**0.45**	1.52%	4.55%	16.67%	18.18%	25.76%	33.33%	"Take 5"	**0.37**	6.14%	7.02%	12.28%	16.67%	28.07%	29.82%
6, 4	"Take tax transfer 6"	**-0.16**	9.09%	23.48%	36.36%	13.64%	13.64%	3.79%	"Take 6"	**-0.29**	13.16%	24.56%	41.23%	14.91%	4.39%	1.75%
7, 3	"Take tax transfer 7"	**-0.35**	20.45%	37.88%	19.70%	8.33%	7.58%	6.06%	"Take 7"	**-0.51**	21.93%	47.37%	23.68%	2.63%	2.63%	1.75%
8, 2	"Take tax transfer 8"	**-0.45**	34.85%	33.33%	10.61%	7.58%	9.09%	4.55%	"Take 8"	**-0.65**	40.35%	43.86%	9.65%	0.88%	3.51%	1.75%
9, 1	"Take tax transfer 9"	**-0.54**	48.48%	24.24%	9.09%	3.79%	11.36%	3.03%	"Take 9"	**-0.74**	59.65%	28.95%	5.26%	1.75%	1.75%	2.63%
10, 0	"Take tax transfer 10"	**-0.61**	61.36%	16.67%	4.55%	3.79%	8.33%	5.30%	"Take 10"	**-0.71**	64.91%	21.05%	5.26%	0.00%	1.75%	7.02%
Panel B: Endowment - 5 Tokens																
Final allocation ("Keep")	Tax framed Democratic norm ratings								Neutrally framed Democratic norm ratings							
	Action	**Mean**	- - -	- -	-	+	+ +	+ + +	**Action**	**Mean**	- - -	- -	-	+	+ +	+ + +
0, 10	"Make tax transfer 5"	**-0.3**	40.91%	15.91%	9.85%	7.58%	12.88%	12.88%	"Give 5"	**0.47**	8.77%	7.89%	3.51%	15.79%	14.04%	50.00%
1, 9	"Make tax transfer 4"	**-0.22**	31.06%	20.45%	9.09%	13.64%	12.12%	13.64%	"Give 4"	**0.53**	1.75%	10.53%	7.89%	12.28%	17.54%	50.00%
2, 8	"Make tax transfer 3"	**-0.09**	15.91%	23.48%	18.94%	13.64%	15.91%	12.12%	"Give 3"	**0.58**	0.88%	6.14%	7.89%	14.91%	23.68%	46.49%
3, 7	"Make tax transfer 2"	**0.09**	6.82%	16.67%	19.70%	22.73%	21.97%	12.12%	"Give 2"	**0.58**	0.88%	2.63%	6.14%	18.42%	35.09%	36.84%
4, 6	"Make tax transfer 1"	**0.22**	4.55%	9.09%	16.67%	31.06%	24.24%	14.39%	"Give 1"	**0.63**	0.00%	0.88%	3.51%	16.67%	44.74%	34.21%
5, 5	"Do not involve government"	**0.67**	2.27%	3.03%	6.82%	6.82%	24.24%	56.82%	"Do nothing"	**0.68**	1.75%	0.88%	3.51%	6.14%	44.74%	42.98%
6, 4	"Take tax transfer 1"	**-0.15**	15.15%	17.42%	32.58%	17.42%	10.61%	6.82%	"Take 1"	**-0.18**	8.77%	18.42%	48.25%	11.40%	9.65%	3.51%
7, 3	"Take tax transfer 2"	**-0.4**	21.21%	37.12%	23.48%	8.33%	7.58%	2.27%	"Take 2"	**-0.43**	16.67%	43.86%	27.19%	7.02%	3.51%	1.75%
8, 2	"Take tax transfer 3"	**-0.48**	34.09%	35.61%	12.88%	5.30%	6.82%	5.30%	"Take 3"	**-0.57**	30.70%	46.49%	15.79%	1.75%	2.63%	2.63%
9, 1	"Take tax transfer 4"	**-0.54**	52.27%	18.94%	9.85%	6.06%	6.82%	6.06%	"Take 4"	**-0.69**	49.12%	38.60%	5.26%	2.63%	1.75%	2.63%
10, 0	"Take tax transfer 5"	**-0.58**	61.36%	12.88%	4.55%	4.55%	13.64%	3.03%	"Take 5"	**-0.72**	71.05%	15.79%	0.88%	1.75%	5.26%	5.26%
Panel C: Endowment - 10 Tokens																
Final allocation ("Keep")	Tax framed Democratic norm ratings								Neutrally framed Democratic norm ratings							
	Action	**Mean**	- - -	- -	-	+	+ +	+ + +	**Action**	**Mean**	- - -	- -	-	+	+ +	+ + +
0, 10	"Make tax transfer 10"	**-0.26**	43.94%	11.36%	8.33%	7.58%	9.09%	19.70%	"Give 10"	**0.44**	11.40%	7.89%	6.14%	13.16%	7.89%	53.51%
1, 9	"Make tax transfer 9"	**-0.15**	31.06%	18.18%	9.85%	9.85%	10.61%	20.45%	"Give 9"	**0.49**	6.14%	10.53%	7.02%	11.40%	11.40%	53.51%
2, 8	"Make tax transfer 8"	**-0.03**	15.15%	27.27%	10.61%	13.64%	13.64%	19.70%	"Give 8"	**0.55**	2.63%	7.02%	10.53%	11.40%	17.54%	50.88%
3, 7	"Make tax transfer 7"	**0.14**	6.82%	13.64%	27.27%	12.88%	19.70%	19.70%	"Give 7"	**0.64**	0.88%	5.26%	6.14%	14.04%	19.30%	54.39%
4, 6	"Make tax transfer 6"	**0.27**	2.27%	7.58%	24.24%	20.45%	26.52%	18.94%	"Give 6"	**0.64**	0.00%	3.51%	5.26%	12.28%	35.09%	43.86%
5, 5	"Make tax transfer 5"	**0.63**	0.76%	3.03%	6.82%	13.64%	28.79%	46.97%	"Give 5"	**0.78**	0.88%	0.00%	1.75%	6.14%	33.33%	57.89%
6, 4	"Make tax transfer 4"	**0.27**	0.76%	5.30%	21.97%	30.30%	30.30%	11.36%	"Give 4"	**0.21**	0.00%	7.02%	28.07%	27.19%	29.82%	7.89%
7, 3	"Make tax transfer 3"	**-0.01**	2.27%	26.52%	24.24%	21.97%	18.94%	6.06%	"Give 3"	**-0.03**	5.26%	21.05%	30.70%	16.67%	21.05%	5.26%
8, 2	"Make tax transfer 2"	**-0.2**	13.64%	34.09%	15.15%	18.18%	13.64%	5.30%	"Give 2"	**-0.21**	8.77%	40.35%	17.54%	15.79%	13.16%	4.39%
9, 1	"Make tax transfer 1"	**-0.35**	31.82%	25.76%	13.64%	12.12%	9.09%	7.58%	"Give 1"	**-0.29**	20.18%	35.09%	14.04%	13.16%	13.16%	4.39%
10, 0	"Do not involve government"	**-0.52**	48.48%	18.94%	12.12%	8.33%	8.33%	3.79%	"Do nothing"	**-0.41**	37.72%	27.19%	10.53%	7.89%	8.77%	7.89%

Notes: 1. Significant at the *** 1 percent, ** 5 percent, and * 10 percent levels.

2. Responses are: "very socially inappropriate" (- - -), "socially inappropriate (- -), "somewhat socially inappropriate" (-), "somewhat socially appropriate" (+), "socially appropriate" (+ +), and "very socially appropriate" (+ + +). Modal responses are shaded in gray. To construct the mean ratings, we converted responses into numerical scores ("very socially inappropriate" = -1, "socially inappropriate" = -0.6, "somewhat socially inappropriate" = -0.2, "somewhat socially appropriate" = 0.2, "socially appropriate" = 0.6, and "very socially appropriate" = 1).

Table A.2: Elicited norms for tax-framed and neutrally-framed Republicans

Panel A: Endowment - 0 Tokens

Final allocation ("Keep")	Tax framed Republican norm ratings								Neutrally framed Republican norm ratings							
	Action	**Mean**	- - -	- -	-	+	+ +	+ + +	**Action**	**Mean**	- - -	- -	-	+	+ +	+ + +
0, 10	"Do not involve government"	**0.16**	27.94%	7.35%	5.88%	5.88%	11.76%	41.18%	"Do nothing"	**0.45**	16.92%	3.08%	3.08%	7.69%	15.38%	53.85%
1, 9	"Take tax transfer 1"	**0.06**	19.12%	16.18%	10.29%	16.18%	10.29%	27.94%	"Take 1"	**0.17**	10.77%	12.31%	18.46%	13.85%	21.54%	23.08%
2, 8	"Take tax transfer 2"	**0.06**	11.76%	25.00%	11.76%	5.88%	29.41%	16.18%	"Take 2"	**0.06**	13.85%	16.92%	15.38%	20.00%	10.77%	23.08%
3, 7	"Take tax transfer 3"	**0.1**	11.76%	13.24%	16.18%	20.59%	23.53%	14.71%	"Take 3"	**0.07**	12.31%	15.38%	12.31%	27.69%	16.92%	15.38%
4, 6	"Take tax transfer 4"	**0.05**	13.24%	19.12%	13.24%	16.18%	23.53%	14.71%	"Take 4"	**0.08**	12.31%	12.31%	23.08%	9.23%	32.31%	10.77%
5, 5	"Take tax transfer 5"	**0.1**	16.18%	14.71%	14.71%	13.24%	14.71%	26.47%	"Take 5"	**0.15**	10.77%	16.92%	15.38%	12.31%	20.00%	24.62%
6, 4	"Take tax transfer 6"	**-0.34**	23.53%	25.00%	27.94%	11.76%	10.29%	1.47%	"Take 6"	**-0.29**	18.46%	29.23%	26.15%	10.77%	12.31%	3.08%
7, 3	"Take tax transfer 7"	**-0.54**	33.82%	38.24%	11.76%	11.76%	2.94%	1.47%	"Take 7"	**-0.51**	26.15%	47.69%	7.69%	13.85%	4.62%	0.00%
8, 2	"Take tax transfer 8"	**-0.62**	50.00%	27.94%	8.82%	4.41%	7.35%	1.47%	"Take 8"	**-0.57**	35.38%	44.62%	4.62%	9.23%	4.62%	1.54%
9, 1	"Take tax transfer 9"	**-0.68**	61.76%	20.59%	2.94%	8.82%	2.94%	2.94%	"Take 9"	**-0.61**	55.38%	21.54%	4.62%	9.23%	6.15%	3.08%
10, 0	"Take tax transfer 10"	**-0.76**	75.00%	11.76%	4.41%	0.00%	5.88%	2.94%	"Take 10"	**-0.67**	64.62%	18.46%	4.62%	1.54%	4.62%	6.15%

Panel B: Endowment - 5 Tokens

Final allocation ("Keep")	Tax framed Republican norm ratings								Neutrally framed Republican norm ratings							
	Action	**Mean**	- - -	- -	-	+	+ +	+ + +	**Action**	**Mean**	- - -	- -	-	+	+ +	+ + +
0, 10	"Make tax transfer 5"	**-0.52**	64.71%	7.35%	4.41%	4.41%	4.41%	14.71%	"Give 5"	**0.35**	15.38%	3.08%	12.31%	9.23%	16.92%	43.08%
1, 9	"Make tax transfer 4"	**-0.46**	47.06%	22.06%	7.35%	7.35%	5.88%	10.29%	"Give 4"	**0.40**	7.69%	10.77%	9.23%	7.69%	24.62%	40.00%
2, 8	"Make tax transfer 3"	**-0.32**	29.41%	27.94%	14.71%	8.82%	8.82%	10.29%	"Give 3"	**0.40**	1.54%	12.31%	12.31%	15.38%	24.62%	33.85%
3, 7	"Make tax transfer 2"	**-0.11**	11.76%	27.94%	17.65%	17.65%	17.65%	7.35%	"Give 2"	**0.43**	1.54%	10.77%	13.85%	9.23%	30.77%	33.85%
4, 6	"Make tax transfer 1"	**0.16**	5.88%	10.29%	25.00%	19.12%	26.47%	13.24%	"Give 1"	**0.48**	3.08%	3.08%	6.15%	21.54%	40.00%	26.15%
5, 5	"Do not involve government"	**0.81**	2.94%	0.00%	2.94%	7.35%	8.82%	77.94%	"Do nothing"	**0.61**	4.62%	1.54%	3.08%	9.23%	41.54%	40.00%
6, 4	"Take tax transfer 1"	**-0.14**	11.76%	25.00%	30.88%	10.29%	11.76%	10.29%	"Take 1"	**-0.22**	12.31%	24.62%	33.85%	16.92%	9.23%	3.08%
7, 3	"Take tax transfer 2"	**-0.45**	27.94%	36.76%	17.65%	5.88%	10.29%	1.47%	"Take 2"	**-0.37**	18.46%	38.46%	26.15%	3.08%	10.77%	3.08%
8, 2	"Take tax transfer 3"	**-0.54**	42.65%	32.35%	10.29%	1.47%	7.35%	5.88%	"Take 3"	**-0.53**	33.85%	38.46%	13.85%	4.62%	7.69%	1.54%
9, 1	"Take tax transfer 4"	**-0.59**	54.41%	25.00%	4.41%	4.41%	4.41%	7.35%	"Take 4"	**-0.66**	46.15%	38.46%	6.15%	3.08%	6.15%	0.00%
10, 0	"Take tax transfer 5"	**-0.68**	72.06%	7.35%	5.88%	5.88%	0.00%	8.82%	"Take 5"	**-0.75**	61.54%	27.69%	3.08%	3.08%	3.08%	1.54%

Panel C: Endowment - 10 Tokens

Final allocation ("Keep")	Tax framed Republican norm ratings								Neutrally framed Republican norm ratings							
	Action	**Mean**	- - -	- -	-	+	+ +	+ + +	**Action**	**Mean**	- - -	- -	-	+	+ +	+ + +
0, 10	"Make tax transfer 10"	**-0.55**	66.18%	8.82%	4.41%	2.94%	1.47%	16.18%	"Give 10"	**0.24**	13.85%	7.69%	18.46%	12.31%	10.77%	36.92%
1, 9	"Make tax transfer 9"	**-0.43**	51.47%	17.65%	5.88%	2.94%	5.88%	16.18%	"Give 9"	**0.27**	10.77%	12.31%	9.23%	18.46%	13.85%	35.38%
2, 8	"Make tax transfer 8"	**-0.35**	36.76%	26.47%	8.82%	7.35%	5.88%	14.71%	"Give 8"	**0.38**	3.08%	12.31%	15.38%	10.77%	23.08%	35.38%
3, 7	"Make tax transfer 7"	**-0.14**	25.00%	20.59%	10.29%	19.12%	8.82%	16.18%	"Give 7"	**0.39**	3.08%	7.69%	16.92%	18.46%	18.46%	35.38%
4, 6	"Make tax transfer 6"	**-0.07**	19.12%	22.06%	13.24%	11.76%	20.59%	13.24%	"Give 6"	**0.49**	3.08%	1.54%	12.31%	18.46%	32.31%	32.31%
5, 5	"Make tax transfer 5"	**0.25**	16.18%	10.29%	13.24%	7.35%	11.76%	41.18%	"Give 5"	**0.76**	0.00%	1.54%	3.08%	9.23%	26.15%	60.00%
6, 4	"Make tax transfer 4"	**-0.03**	7.35%	20.59%	25.00%	22.06%	19.12%	5.88%	"Give 4"	**0.35**	0.00%	7.69%	16.92%	23.08%	35.38%	16.92%
7, 3	"Make tax transfer 3"	**-0.09**	8.82%	17.65%	29.41%	30.88%	8.82%	4.41%	"Give 3"	**0.13**	0.00%	16.92%	29.23%	16.92%	27.69%	9.23%
8, 2	"Make tax transfer 2"	**-0.17**	13.24%	26.47%	22.06%	19.12%	16.18%	2.94%	"Give 2"	**0.00**	7.69%	26.15%	20.00%	12.31%	23.08%	10.77%
9, 1	"Make tax transfer 1"	**-0.15**	25.00%	23.53%	10.29%	11.76%	13.24%	16.18%	"Give 1"	**-0.06**	12.31%	29.23%	12.31%	13.85%	23.08%	9.23%
10, 0	"Do not involve government"	**-0.13**	36.76%	14.71%	5.88%	5.88%	10.29%	26.47%	"Do nothing"	**-0.18**	27.69%	20.00%	9.23%	13.85%	21.54%	7.69%

Notes: 1. Significant at the *** 1 percent, ** 5 percent, and * 10 percent levels.

2. Responses are: "very socially inappropriate" (- - -), "socially inappropriate (- -), "somewhat socially inappropriate" (-), "somewhat socially appropriate" (+), "socially appropriate" (+ +), and "very socially appropriate" (+ + +). Modal responses are shaded in gray. To construct the mean ratings, we converted responses into numerical scores ("very socially inappropriate" = -1, "socially inappropriate" = -0.6, "somewhat socially inappropriate" = -0.2, "somewhat socially appropriate" = 0.2, "socially appropriate" = 0.6, and "very socially appropriate" = 1).

Table A.3: Wilcoxon signed-rank tests testing equality of norm ratings across endowments

Keep	Endowments 0 vs. 5	Endowments 0 vs. 10	Endowments 5 vs. 10
0	1.398	2.922	1.754
1	-3.690***	-2.201	2.233
2	-4.721***	-4.314***	1.105
3	-5.010***	-6.087***	-0.648
4	-6.500***	-6.958***	-0.913
5	-6.297***	-8.130***	-3.422**
6	-2.246	-8.926***	-7.663***
7	-3.319**	-9.234***	-7.591***
8	-2.223	-8.450***	-7.459***
9	-1.278	-8.181***	-8.436***
10	1.637	-5.999***	-7.591***

Significant at the *** 1 percent, ** 5 percent, and * 10 percent levels with Bonferroni correction; all two-tailed.

Table A.4: Mann-Whitney U tests testing neutrally framed Democratic and Republican norm ratings

Keep	Endowment 0	Endowment 5	Endowment 10
0	0.504	1.062	1.967
1	1.801	1.298	2.218
2	2.478	1.979	1.979
3	2.247	1.270	2.887
4	2.105	1.900	2.061
5	1.988	0.761	0.031
6	0.442	0.572	-1.985
7	0.382	-0.472	-1.877
8	-0.893	-0.228	-2.232
9	-1.107	-0.500	-2.284
10	-0.152	-0.933	-2.117

Significant at the *** 1 percent, ** 5 percent, and * 10 percent levels with Bonferroni correction; all two-tailed.

Table A.5: Mann-Whitney U tests testing the effect of frame on norm ratings by identity

Panel A: Tax framed and neutrally framed Democrats			
Keep	Endowment 0	Endowment 5	Endowment 10
0	5.775***	7.506***	6.500***
1	2.811	7.742***	6.352***
2	2.052	7.709***	6.343***
3	1.090	6.681***	6.440***
4	0.773	6.471***	5.679***
5	-0.743	-1.191	2.488
6	-1.806	-0.302	-1.084
7	-1.908	-0.189	-0.154
8	-2.194	-0.617	-0.034
9	-2.459	-0.873	1.078
10	-0.867	-1.755	1.442
Panel B: Tax framed and neutrally framed Republicans			
Keep	Endowment 0	Endowment 5	Endowment 10
0	1.870	5.733***	5.646***
1	0.716	6.16***	5.049***
2	0.087	5.74***	5.487***
3	-0.288	4.881***	4.241***
4	0.264	3.473**	4.595***
5	0.355	-3.872***	3.606**
6	0.504	-0.544	3.931***
7	0.541	1.111	2.321
8	1.129	0.706	1.523
9	0.826	0.331	0.841
10	1.249	0.637	-0.268

Significant at the *** 1 percent, ** 5 percent, and * 10 percent levels with Bonferroni correction; all two-tailed.

A.2 Additional tests of the initial endowments

We also find that the initial endowment affects dictator choice even in the neutral frame. According to an OLS regression, for each extra token initially endowed to the dictator, the dictator keeps an additional 0.059 tokens ($p < 0.01$). As noted in the literature review, this result contrasts with the majority of prior work examining the effect of the initial endowment on dictator behavior.

Several differences between our design and previous work could account for these conflicting results. First, we include behavior for all integer initial endowments between 0 and 10, whereas most of the previous work only includes the extreme cases where the dictator starts with either all or none of the endowment (*Dreber et al.*, 2013; *Grossman and Eckel*, 2012, 2015; *Halvorsen*, 2015; *Hauge et al.*, 2016; *Goerg et al.*, 2017). Second, we have a within-subjects design where all subjects make a sequence of dictator decisions for all possible initial endowment distributions. Thus, our subjects experience changes in the initial endowment.

Due to our experimental design, we are able to restrict our data to examine which of these two differences, changes in initial endowments or extreme initial endowments, are more responsible for the effects we see in this paper. First, we perform a within-subjects analysis of the extreme distributions. To do so, we take all the data from the neutral frame and restrict our attention to the dictator allocations when the initial endowments are (10,0) or (0,10). Note that this means that we will have two behavioral observations for each dictator. When we run a Wilcoxon signed-rank test, we find that there are significant differences in how much the dictator keeps when they start with 0 or 10 tokens ($p < 0.01$). This test provides an opportunity to more directly compare our results to those obtained in *Visser and Roelofs* (2011) and *Korenok et al.* (2014). Both of those papers use a within-subjects design and vary the distribution of the initial endowment between the dictator and recipient. Both also find that when the recipient starts with all of the initial endowment, the final payoff

for the recipient is higher – i.e. the dictator tends not to take as much for herself. Thus, all three papers (ours and these) find a significant effect of initial endowments on dictator behavior when dictators are exposed to a sequence of changing initial endowments.

Next, we perform a between-subjects analysis on the non-extreme distribution of (5,5). Specifically, we restrict our data in the neutral frame to the first choice that each dictator makes. In our experiment, this first choice was under the initial endowment distribution of (10,0), (5,5), or (0,10). This restriction removes any impact of prior exposure to other initial endowments, giving us a between-subjects design. We run the Kruskal-Wallis test on this data. We find that there is no difference in the distributions of final allocations ($p = 0.413$).

Restricting our data to only the dictator's first choice also provides us an opportunity to more directly compare our results to those obtained in a working paper by *Grossman and Eckel* (2012) as well the results of *Krupka and Weber* (2013a). In *Grossman and Eckel* (2012), the authors run three between-subjects treatments: those with initial endowments of ($20, $0), ($10, $10), and ($0, $20). They find no difference between the final amounts donated to a charity when the initial endowments are ($20,$0) and ($0,$20), but do find a significant difference when the initial endowments are ($10,$10). However, when we run the same analysis (a Fligner-Policello Robust Rank Order Test) on our data, we find no significant differences between any of these three treatments (Endowment 0 vs. 10, $p = 0.276$; Endowment 5 vs. 10, $p = 0.968$; Endowment 0 vs. 5, $p = 0.220$).

We also present the non-parametric equivalent to the regression in the paper to make the result comparable to the previous results. For this test, we restrict our attention to the neutral frame and to those subjects whose first choice was under either the (10,0) or (0,10) initial endowment. With these restrictions, we run a Mann-Whitney U test and find no significant effect ($p = 0.233$). The conclusion that our

data support is that a sequence of changes in the initial endowment significantly impacts dictators' final allocation decisions. The support for this is most readily seen in the first Wilcoxon signed-rank test testing the differences of the extreme distribution (between-subjects, endowment 10 vs. endowment 0). The interpretation we offer is that the changes in initial endowments act as a procedural frame (*Larrick and Blount*, 1997; *Kahneman*, 2000). This is consistent with reference dependence-changes from an initial point are salient.

Our combination of results can offer a way to think about the conflicting results in the literature. The prior work focuses on what happens to dictator allocations when they are exposed to one initial endowment. The treatments then vary the initial endowment using a between-subjects design allocating the entire endowment to either the dictator or the recipient. This prior work tends to find that initial endowment distributions do not affect final allocations. Our work cannot speak to other nuances in the literature that may be contributing to differences. Specifically, there is an interesting observation to be made when examining *Krupka and Weber* (2013a) and *Grossman and Eckel* (2012). In those papers, the designs are between-subjects and the baseline is a standard dictator game (where the initial endowment rests with the dictator). The treatment is a dictator game where 50% of the initial endowment rests with the dictator and 50% rests with the recipient. Comparing their baseline to their treatment, they both find significant differences. Though we too are able to restrict our analysis to a between-subjects comparison of the initial endowments of (10,0) or (5,5), we find no effect on final allocations. However, we struggle to perfectly identify why we obtain no effect as we cannot rule out subject pool differences (e.g. we use MTurkers) as well as differences in payoff size (our payoffs range between \$0 and \$1). There is evidence, for example, that stake size matters substantially for these framing effects (*Leibbrandt et al.*, 2015).

A.3 Experimental Instructions

A.3.1 Choice experiments

In the *choice experiments*, participants in both tax- and neutrally-framed treatments read the same introduction. These participants also complete the same 10-item questionnaires and demographic questionnaire shown below. The only difference between the treatments is in the framing.

A.3.1.1 Tax-framed

Introduction

[Overview of Tasks]

This is a study in decision making that has three parts. You will earn a 50 cent base pay for completing the study.

In the first part, we will ask you to tell us what you think about various images.

In the second part, you will have a chance to earn a bonus. Your earnings for the second part will be in tokens, which will be converted to money. Every 10 tokens you earn is worth $1 to you. Your earnings will depend on the decisions you make and on the decision that the other worker you paired with will make.

In the final third part, we will ask you to tell us about yourself.

You will be paid the base plus the bonus within 3 days after you complete this task.

Note: If you are using Internet Explorer you will not be able to complete the survey. Please try using Safari, Firefox, or Chrome

10-item questionnaire

[Tell Us What You Think]

You will now be shown several pairs of pictures of people. Please indicate which person in each pair you find more attractive.

[Tell Us What You Think]

Please indicate which person in each pair you find more attractive.

Fig. A.1 is an example of what these pairs of pictures of people look like. They make selection between 5 different pairs of images.

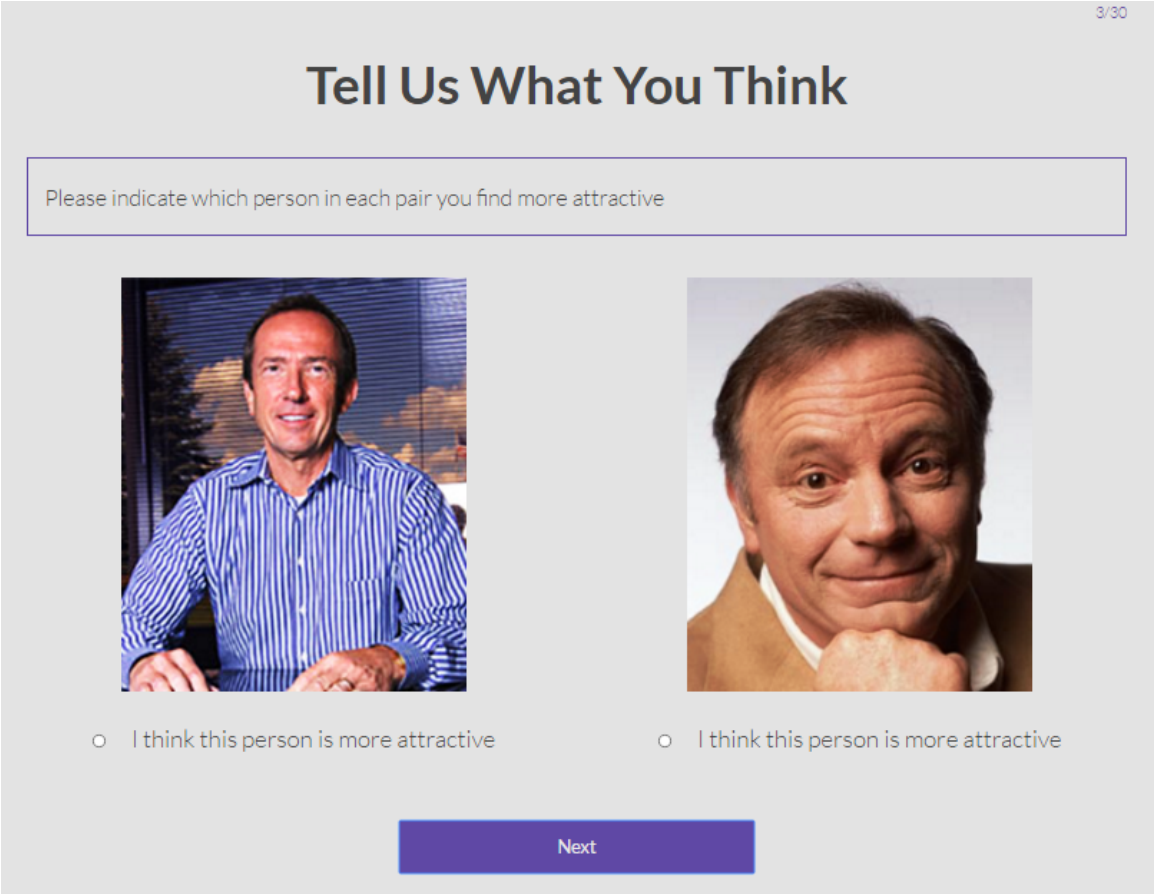

Figure A.1: Screen shot of one of the person selection task in the 10-item questionnaire. Participants in all treatments make a selection for 5 similar sets of images of individuals.

[Tell Us What You Think]

On the next screen you will see two images of people waiting in lines. Please indicate which line you think is the longest.

[Tell Us What You Think]

Please indicate which line you think is the longest.

Fig. A.2 is an example of what this pair of images of lines looks like. They make this selection once.

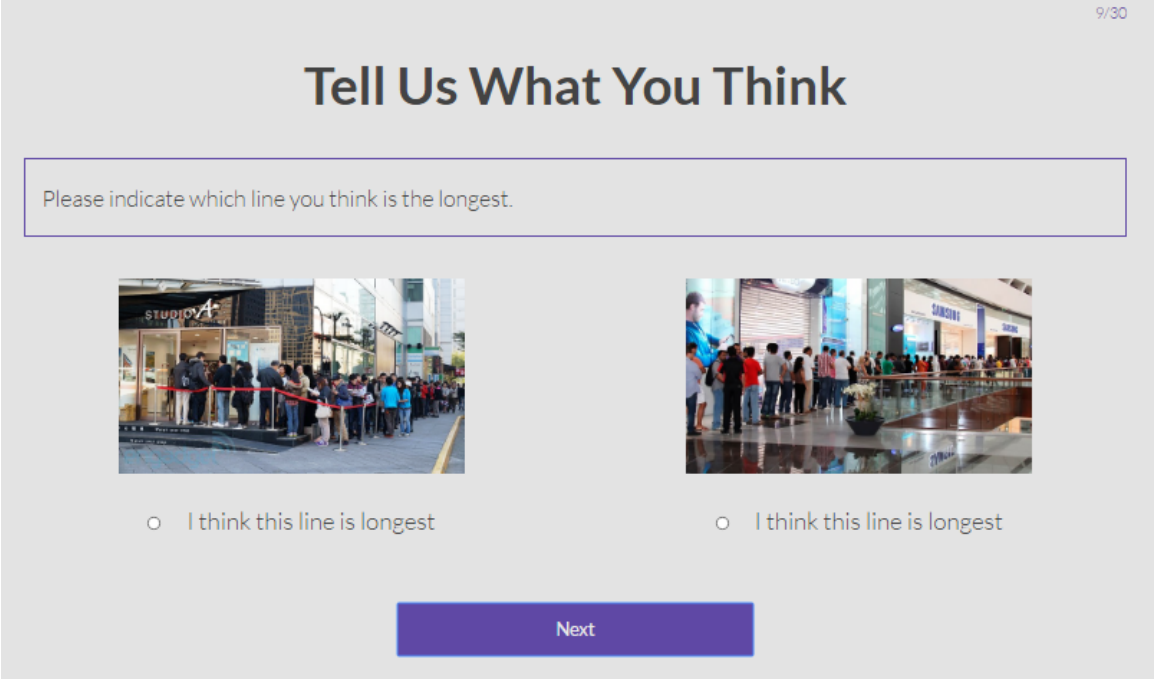

Figure A.2: Screen shot of the line length selection task in the 10-item questionnaire. Participants in all treatments see the same pair of images and make a selection for this pair.

[Tell Us What You Think]

You will now be shown several states. For each state, please answer the following question: What was the state's average temperature in 2013?

[Tell Us What You Think]

What was the state's average temperature in 2013?

Fig. A.3 is an example of what these temperature selection questions look like. They make this selection for 4 different states.

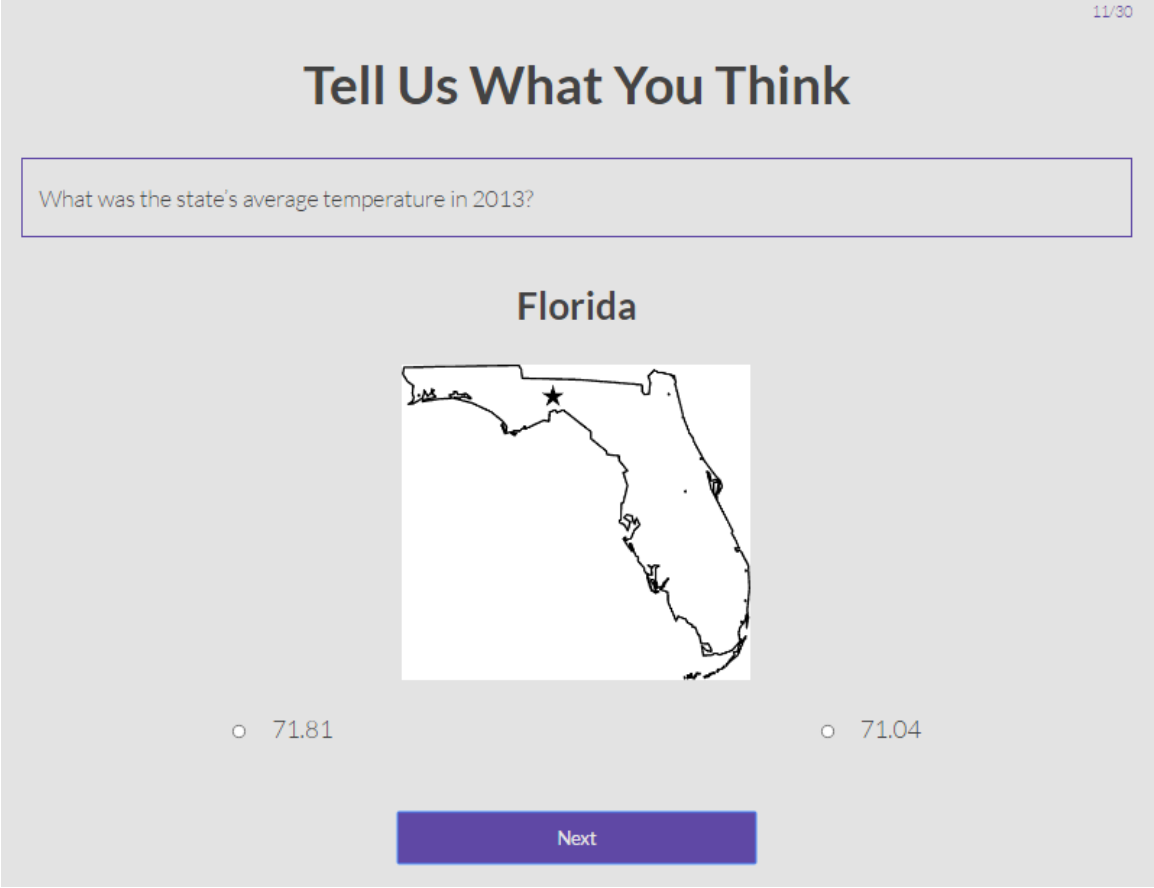

Figure A.3: Screen shot of one the states in the state temperature selection task in the 10-item questionnaire. Participants in all treatments make a selection for 4 different states.

Tax-framed dictator games

[Bonus Task]

For the following task, you will be randomly paired with another person, whom we will call your match. The match will be randomly selected from the other workers.

[Bonus Task]

In our economy one way the government uses taxes is to generate revenue from its citizens' earnings to redistribute wealth. The government's role in redistributing this wealth can be large or small. Sometimes people have a lot of wealth in our economy and sometimes people have little wealth in our economy.

You have the opportunity to tell the government if it should get involved in wealth redistribution between you and your match and, if so, how large or small the redistribution should be. If your decision is selected for payment, it will determine how many tokens each person gets paid in this task.

[Bonus Task]

When you and your match have entered all of your decisions, we will then randomly pick one of the decisions from the set that you and your match made. The selected decision will determine the final token split between you and your match and will be paid out to you as a bonus for this task.

[Bonus Task]

In this economy your wealth is X tokens and your match's wealth is Y tokens.

Use the slider to indicate whether you want the government involved and how large or small the redistribution should be.

Make a decision by moving the slider.

To confirm, your post-tax wealth for this decision is W tokens.

To confirm, your match's post-tax wealth for this decision is V tokens.

Fig. A.4 is a screen shot of the decision page that participants saw. They may move the slider to indicate their choice. They do this for 11 different endowment

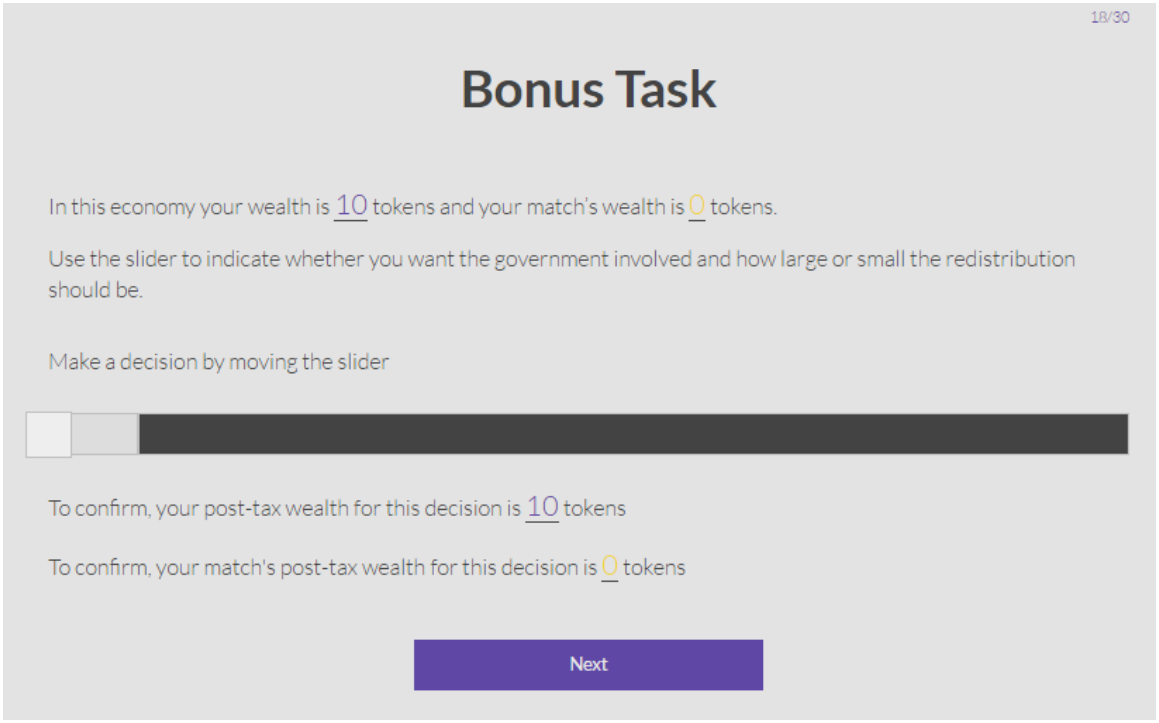

Figure A.4: Screen shot of one of the decision screens for individuals in the tax-framed *choice experiment*. Participants make a choice by moving the slider. The fields update dynamically as they move the slider to reflect the final amount of allocation for themselves and for the other worker. Participants make this selection for endowments 0 to 10.

scenarios, ranging from when they are endowed with all 10 tokens and their match is endowed with 0 tokens down to when they are endowed with 0 tokens and their match is endowed with all 10 tokens.

Demographic questionnaire

[Tell Us About Yourself]

Please complete the following demographic survey. Your responses will not be connected to your worker ID.

[Tell Us About Yourself]

In politics, as of today, do you consider yourself: [A Republican, A Democrat, Leaning more towards the Democratic party, Leaning more towards the Republican party]

What is your age?

What is your gender? [Male, Female]

Which one of the following best describes your racial or ethnic background? [Asian/Pacific Islander, Black, Hispanic/Latino, White, Other]

Have you ever voted in a government election? [Yes/No]

A.3.1.2 Neutrally-framed

Participants in the neutrally-framed treatment complete the same 10-item questionnaire and demographic questions as participants in the tax-framed treatment. However, instead of playing the tax-framed dictator games, participants in this treatment play 11 neutrally-framed dictator games.

Neutrally-framed dictator games

[Bonus Task]

For the following task, you will be randomly paired with another person, whom we will call your match. The match will be randomly selected from the other workers.

[Bonus Task]

You will be shown 11 situations. In each situation, at least one of you will be holding some number of tokens. You will decide whether you would like to give some tokens to your match, take some tokens from your match or do nothing.

[Bonus Task]

When you and your match have entered all of your decisions, we will then randomly pick one of the decisions from the set that you and your match made. The selected decision will determine the final token split between you and your match and will be paid out to you as a bonus for this task.

[Bonus Task]

For this decision you own X tokens and the other person owns Y tokens.

You have the opportunity to take any of the X tokens from the other person. If this decision is selected for payment this will determine how many tokens each person gets.

Use the slider to indicate how many tokens you wish to take or give.

Make a decision by moving the slider.

To confirm, your earnings for this decision will be W tokens.

To confirm, the other person's earnings for this decision will be V tokens.

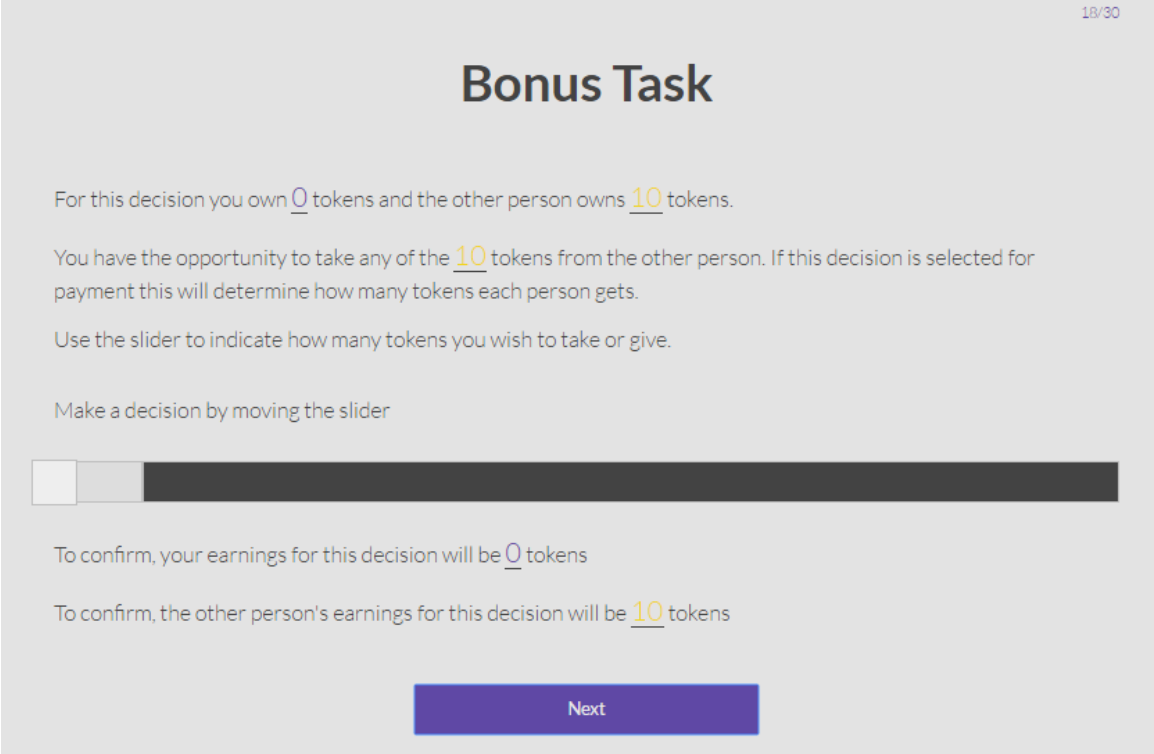

Figure A.5: Screen shot of one of the decision screens for individuals in the neutrally-framed *choice experiment*. Participants make a choice by moving the slider. The fields update dynamically as they move the slider to reflect the final amount of allocation for themselves and for the other worker. Participants make this selection for endowments 0 to 10.

Fig. A.5 is a screen shot of the decision page that participants saw. They may move the slider to indicate their choice. As in the tax-framed treatment, they do this for 11 different endowment scenarios, ranging from when they are endowed with all 10 tokens and their match is endowed with 0 tokens down to when they are endowed with 0 tokens and their match is endowed with all 10 tokens.

A.3.2 Norms elicitation experiments

Participants in the tax- and neutrally-framed treatments read the same introduction. They also complete the same demographics questionnaire and 10-item questionnaire as those participants in the *choice experiment*.

A.3.2.1 Tax-framed

Introduction

> [Overview of Tasks]
>
> This is a study in decision making that has three parts. You will earn a 50 cent base pay for completing the study.
>
> In the first part, we will ask you to tell us about yourself.
>
> In the second part, we will ask you to tell us what you think about various images.
>
> In the third part, you will have a chance to earn a bonus. Your earnings for this part will depend on the decisions you make and on the decisions that the other worker you are paired with makes. You can earn up to $3.30 in bonus pay.
>
> You will be paid the base plus the bonus within 3 days after you complete this task.

Note: If you are using Internet Explorer you will not be able to complete the survey. Please try using Safari, Firefox, or Chrome

[Ready to Start Part One!]

We are ready to start the first part. This is where you tell us about yourself.

Participants then complete the same demographic questionnaire as in the *choice experiment*. Then participants read instructions for the coordination games. The language of these coordination game differ by the frame in which the game is described in.

Explanation of coordination games with tax-framed language

[Explaining How You Will Earn Money In The Bonus Task]

On the next screens you will read about decisions that another Mturker made in a previous Hit. We will call this Mturker "worker A". Worker A is NOT participating today, but made choices in a previous Hit. You will read about the decisions worker A faced and what actions worker A had to choose between.

[Explaining How You Will Earn Money In The Bonus Task]

In our economy one way the government uses taxes is to generate revenue from its citizens' earnings to redistribute wealth. The government's role in redistributing this wealth can be large or small. Sometimes people have a lot of wealth in our economy and sometimes people have little wealth in our economy.

Worker A was randomly paired with another Mturker, called worker B. Worker A faced several different situations in which he or she had the opportunity to tell the government if it should get involved in wealth

redistribution between them and worker B and, if so, how large or small the redistribution should be. Their wealth was represented by tokens, where every 10 tokens was worth $1.

[Explaining How You Will Earn Money In The Bonus Task]

Your job is to rate worker A's wealth redistribution decision based on whether you think the decision was

"socially appropriate"

and

"consistent with what most people who are like you

think that worker A OUGHT to do".

That sounds simple, but it is only half the story!

Specifically, you will only earn the bonus if your "social appropriateness" rating MATCHES the rating of another Mturker working on this HIT today who is like you. We will call this Mturker "your match."

We will match you with another Mturker who is like you. To increase the chances that you earn the bonus, you should try to imagine what your match, who is like you, would say.

Then participants complete the same 10-item questionnaire as those in the *choice experiments*.

Coordination games with tax-framed language

Participants then make rating decisions for a single endowment for each of the possible actions that the dictator could take (e.g. redistributing 0 to 10 tokens). They make the same rating decisions for when the dictator has an endowment of

0, 5, and 10 tokens. Below is an example of what a participant in the tax-framed treatment sees when rating the 11 dictator choices in the scenario where the dictator is endowed with 5 tokens.

[Ready To Start Part Three - The Bonus Task]

We are ready to start part three: This is where you can earn a bonus!

[Bonus Task]

On the next screens you will read about decisions that worker A, an Mturker from another HIT, made. The description will include possible actions available to worker A.

Your task is to rate worker A's wealth redistribution decision based on your guess of whether your MATCH, who is like you, would think the decision was "socially appropriate" and "consistent with what someone who is like you would think worker A OUGHT to do."

Remember that you will only earn the bonus if your "social appropriateness" rating is that same as your MATCH's rating. For each rating that is the same, you will earn 10 cents.

[Bonus Task]

In this economy worker A's wealth was 5 tokens and worker B's wealth was 5 tokens.

Worker A was able to decide whether the government should get involved and how large or small the redistribution should be.

Worker A got the government involved and chose to take a tax transfer of 5 tokens from worker B.

As a result:

Worker A's post-tax wealth for this decision was 10 tokens.

Worker B's post-tax wealth for this decision was 0 tokens.

TASK: Your task is to rate worker A's wealth redistribution decision based on your guess of whether your MATCH would think that the decision is "socially appropriate" and "consistent with what someone who is like you would think worker A OUGHT to do."

I think my MATCH would rate this decision as ["Very socially appropriate," "Socially appropriate," "Somewhat socially appropriate," "Somewhat socially inappropriate," "Socially inappropriate," "Very socially inappropriate."]

Fig. A.6 is a screen shot of the decision page that participants saw. They may select a rating by moving their mouse over the drag down box. Participants rate each of the 11 actions (that would result in Worker A having post-tax wealth of 10 to 0, before moving on to the next endowment and rating the next set of 11 actions for for that endowment. Participants do this for endowments 0, 5 (as in our example), and 10.

A.3.2.2 Neutrally-framed

Participants in the neutrally-framed treatment complete the same 10-item questionnaire and demographic questionnaire as participants in the tax-framed treatment. The only differences in instructions for these participants and their tax-framed counterparts are the explanation of the coordination games and the actual coordination games.

Explanation of coordination games with neutrally-framed language

Instead of the instructions that the tax-framed participants receive earlier, participants in this treatment read the following:

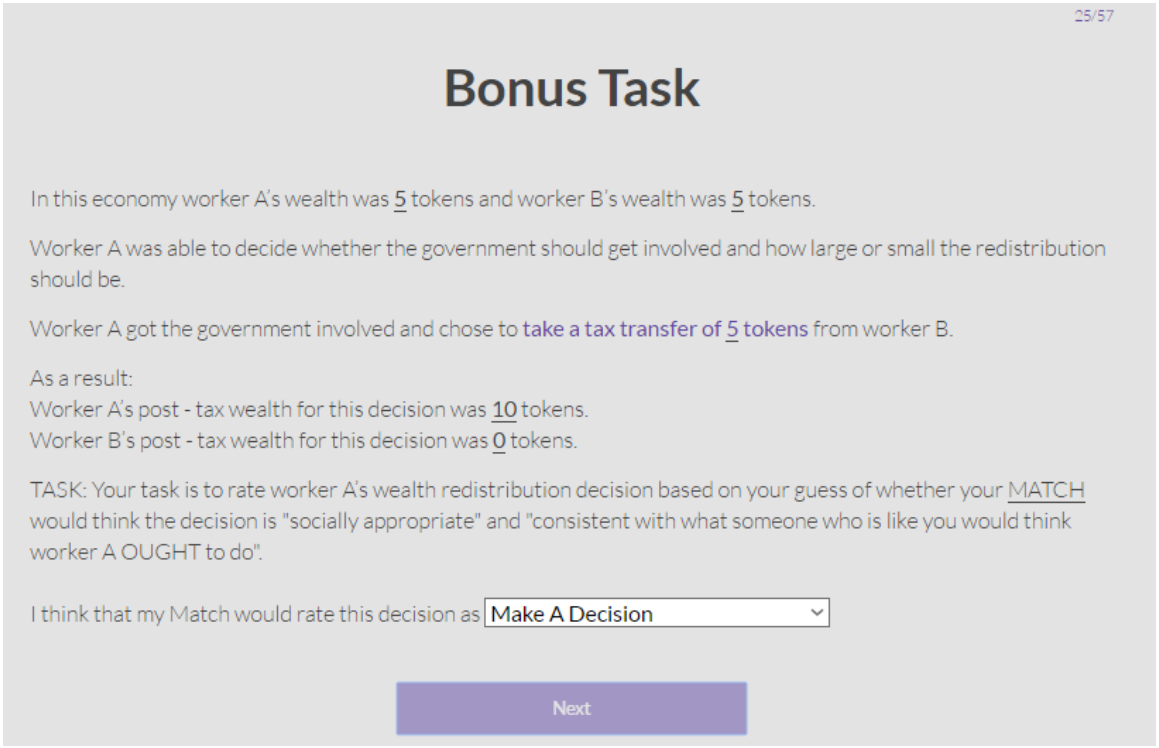

Figure A.6: Screen shot of one of the decision screens for individuals in the tax-framed *norms elicitation experiment*. Participants make a choice by selecting from the drop down box. Participants make this selection for all 11 action (resulting in the participants holding 0 to 10 tokens at the end of the reallocation). They do this for endowments 0, 5, and 10.

[Explaining How You Will Earn Money In The Bonus Task]

On the next screens you will read about decisions that another Mturker made in a previous Hit. We will call this Mturker "worker A." Worker A is NOT participating today, but made choices in a previous Hit. You will read about the decisions worker A faced and what actions worker A had to choose between.

[Explaining How You Will Earn Money In The Bonus Task]

Worker A was randomly paired with another Mturker, called worker B. Worker A faced several different situations in which he or she was holding some number of tokens, where every 10 tokens was worth $1. Worker A then had to decide whether he or she would like to give some tokens to worker A, take some tokens from worker B, or do nothing.

[Explaining How You Will Earn Money In The Bonus Task]

Your job is to rate worker A's decision based on whether you think the decision was

"socially appropriate"

and

"consistent with what most people who are like you think that worker A OUGHT to do".

That sounds simple, but it is only half the story!

Specifically, you will only earn the bonus if your "social appropriateness" rating MATCHES the rating of another Mturker working on this HIT today who is like you. We will call this Mturker "your match."

> We will match you with another Mturker who is like you. To increase the chances that you earn the bonus, you should try to imagine what your match, who is like you, would say.

Then these participants complete the same 10-item questionnaire as those in the *choice experiments*. Instead of the instructions for the tax-framed coordination game, participants in this treatment read about and play neutrally-framed coordination for 3 different endowments, as in the tax-framed treatment. Similar to their counterparts in the tax-framed treatment, these participants make rating decisions for each of the 11 possible actions for when the dictator has an endowment of 0, 5, and 10 tokens.

Below is an example of what a participant in the neutrally-framed treatment sees when rating the 11 dictator choices in the scenario where the dictator is endowed with 0 tokens and the match is endowed with all 10 tokens.

Coordination games with neutrally-framed language

> [Ready To Start Part Three - The Bonus Task!]
>
> We are ready to start part three: This is where you can earn a bonus!
>
> [Bonus Task]
>
> On the next screens you will read about decisions that worker A, an Mturker from another HIT, made. The description will include possible actions available to worker A.
>
> Your task is to rate worker A's decision based on your guess of whether your MATCH, who is like you, would think the decision was "socially appropriate" and "consistent with what someone who is like you would think worker A OUGHT to do".

Remember that you will only earn the bonus if your "social appropriateness" rating is the same as your MATCH's rating. For each rating that is the same, you will earn 10 cents.

[Bonus Task]

For this decision worker A owns 0 tokens and worker B owns 10 tokens.

Worker A had the opportunity to take any amount of worker B's 10 tokens from worker B.

Worker A choose to take 10 tokens from worker B.

As a result:

Worker A's post-earnings for this decision were 10 tokens.

Worker B's post-earnings for this decision were 0 tokens.

TASK: Your task is to rate worker A's wealth redistribution decision based on your guess of whether your MATCH would think that the decision is "socially appropriate" and "consistent with what someone who is like you would think worker A OUGHT to do."

I think my MATCH would rate this decision as ["Very socially appropriate," "Socially appropriate," "Somewhat socially appropriate," "Somewhat socially inappropriate," "Socially inappropriate," "Very socially inappropriate."]

Fig. A.7 is a screen shot of the decision page that participants saw. Similar to the tax-framed condition, participants rate each of the 11 actions (that would result in Worker A having post-earnings of 10 to 0, before moving on to the next endowment and rating the next set of 11 actions for for that endowment. Participants do this for endowments 0, 5 (as in our example), and 10.

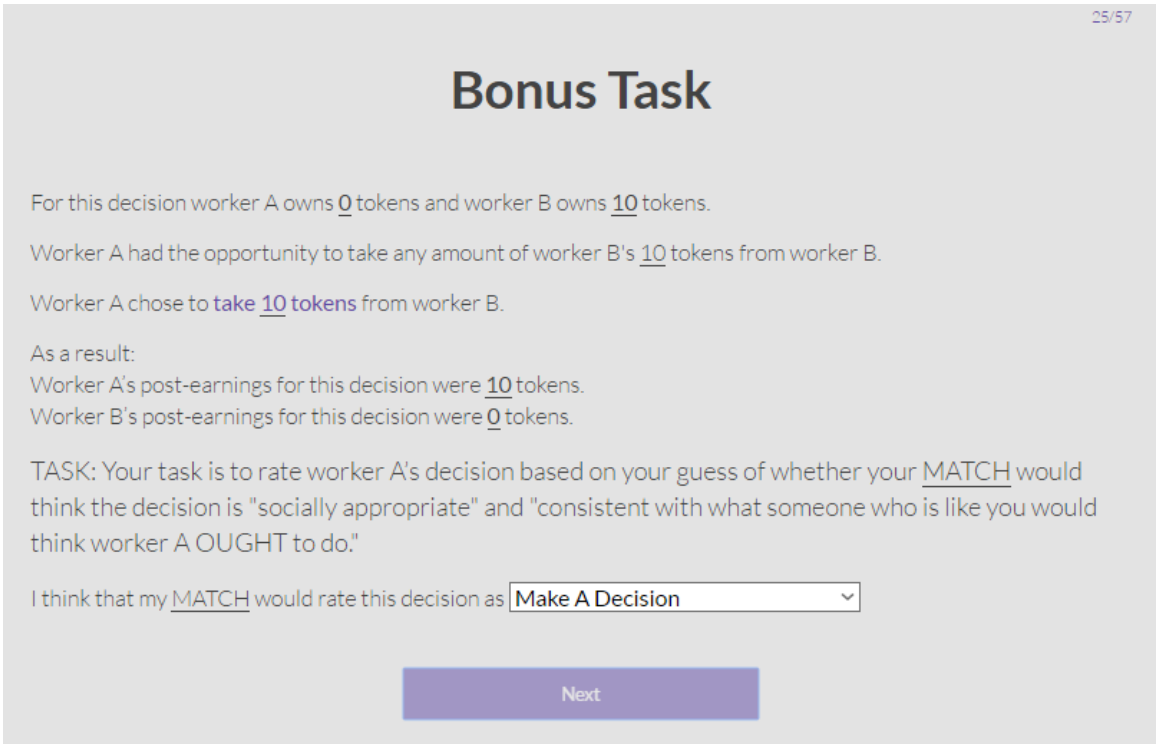

Figure A.7: Screen shot of one of the decision screens for individuals in the neutrally-framed *norms elicitation experiment*. Participants make a choice by selecting from the drop down box. Participants make this selection for all 11 action (resulting in the participants holding 0 to 10 tokens at the end of the reallocation). They do this for endowments 0, 5, and 10.

APPENDIX B

Information Wars

B.1 Additional figure and table of the worker experiment

Table B.1: Logistic regressions of the dummy variable for the hired worker cheating in the bonus task on the hired worker's response.

	(1) A cheated as much as: Dummy variable for worker cheated	(2) A cheated less: Dummy variable for worker cheated	(3) A cheated more: Dummy variable for worker cheated
Sweet	-0.790	0.511	-0.381
	(0.599)	(0.530)	(0.858)
Mean		0.598	1.005
		(1.309)	(1.152)
Constant	0.847*	0.095	0.381
	(0.492)	(0.440)	(0.247)
Observations	55	75	80

Robust standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1

B.2 Additional table of the manager experiment

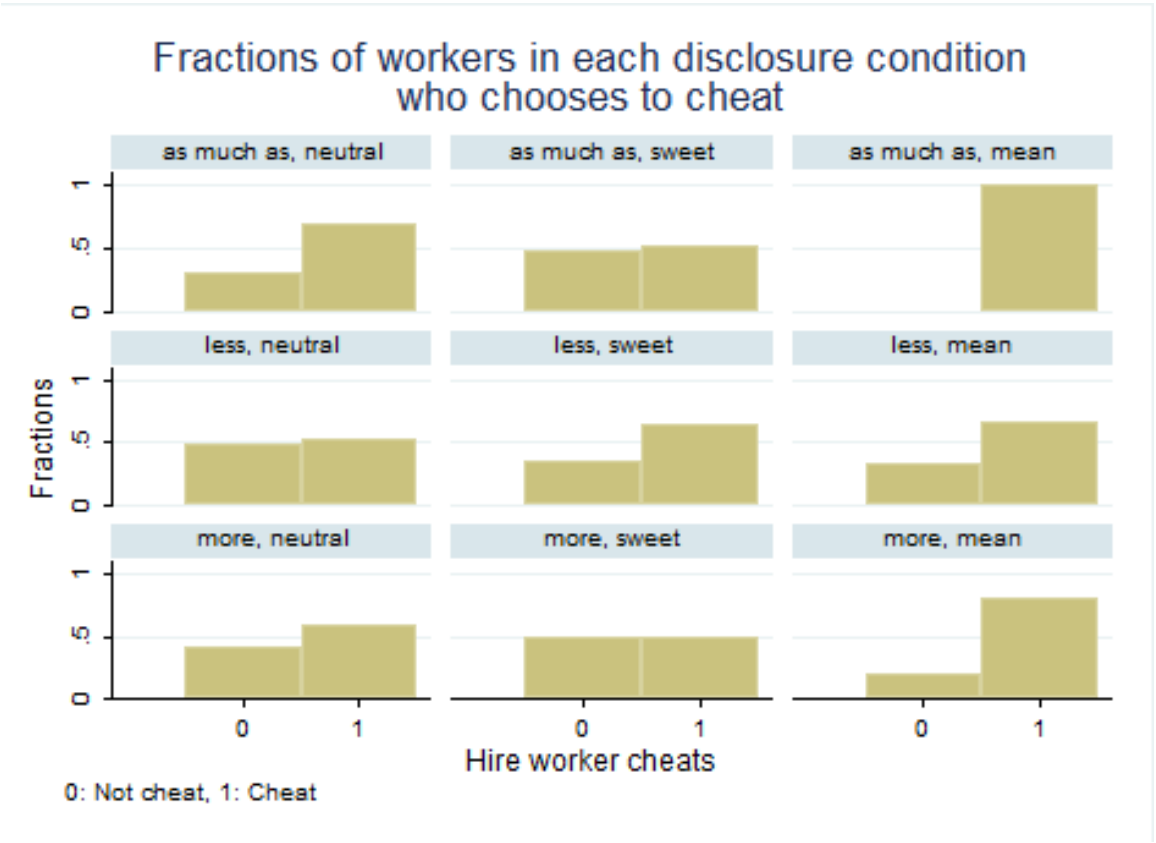

Figure B.1: The fractions of hired workers who choose to cheat in the *workers' bonus task* given the performance disclosure condition that the worker was in and the response that he or she sent.

Table B.2: OLS regression of the reported posterior means on the dummy variables of each of the response conditions, the dummy variables of each of the performance disclosure conditions, and their interaction terms.

	Posterior means after Round 3
Sweet	-0.069
	(0.188)
Mean	0.116
	(0.325)
A cheated less	-0.409**
	(0.194)
A cheated more	0.453**
	(0.208)
Sweet X A cheated less	-0.163
	(0.239)
Sweet X A cheated more	0.158
	(0.25)
Mean X A cheated less	-0.731*
	(0.41)
Mean X A cheated more	-0.004
	(0.407)
Constant	2.291***
	(0.15)
Observations	557
R-squared	0.206

Robust standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1

B.3 Additional results of worker experiment

B.3.0.1 Workers' perceptions of response valence and effectiveness

All workers reported how positive or negative they find each of the responses on a 5-point Likert scale. I convert the 5-points into a number between 1 to 5, where 1 is "very negative," 2 is "somewhat negative," 3 is "neutral," 4 is "somewhat positive," and 5 is "very positive." Figure B.2 presents the average ratings from this question. The "neutral" response is rated as significantly less positive than the "sweet" response (t-test, two-tailed, p-value < 0.01) and more positive than the "mean" response (t-test, two-tailed, p-value < 0.01). The "sweet" response is rated as significantly more positive than the "neutral" response (t-test, two-tailed, p¡0.01) as well as the "mean" response (t-test, two-tailed, p-value < 0.01). The "mean" response is also rated as significantly less positive than either the "neutral" response (t-test, two-tailed, p-value < 0.01) or the "sweet" response (t-test, two-tailed, p-value $p < 0.01$).

Both sets of workers also rated how effective they thought each of the responses was in influencing the manager to hire the sender of that response. They did so on a 4-point Likert scale, where 1 is "very ineffective," 2 is "somewhat ineffective," 3 is "somewhat effective," and 4 is "very effective." Figure B.3 presents the average worker perception of effectiveness for each response. Of the three responses, the "sweet" response is rated as more effective than the "neutral" response and the "mean" response (t-tests, two-tailed, p-value < 0.01 for both). "Neutral" response is considered less effective than the "sweet" response (t-test, two-tailed, p-value < 0.01), but more effective than the "mean" response (t-test, two-tailed, p-value < 0.01). Finally, the "mean" response is considered to be significantly less effective than the "sweet" response and the "mean" response (t-tests, two-tailed, p-value < 0.01).

Next, I look at which response workers believe would be the most effective in persuading a manager. To do this, I include only workers' responses that do not

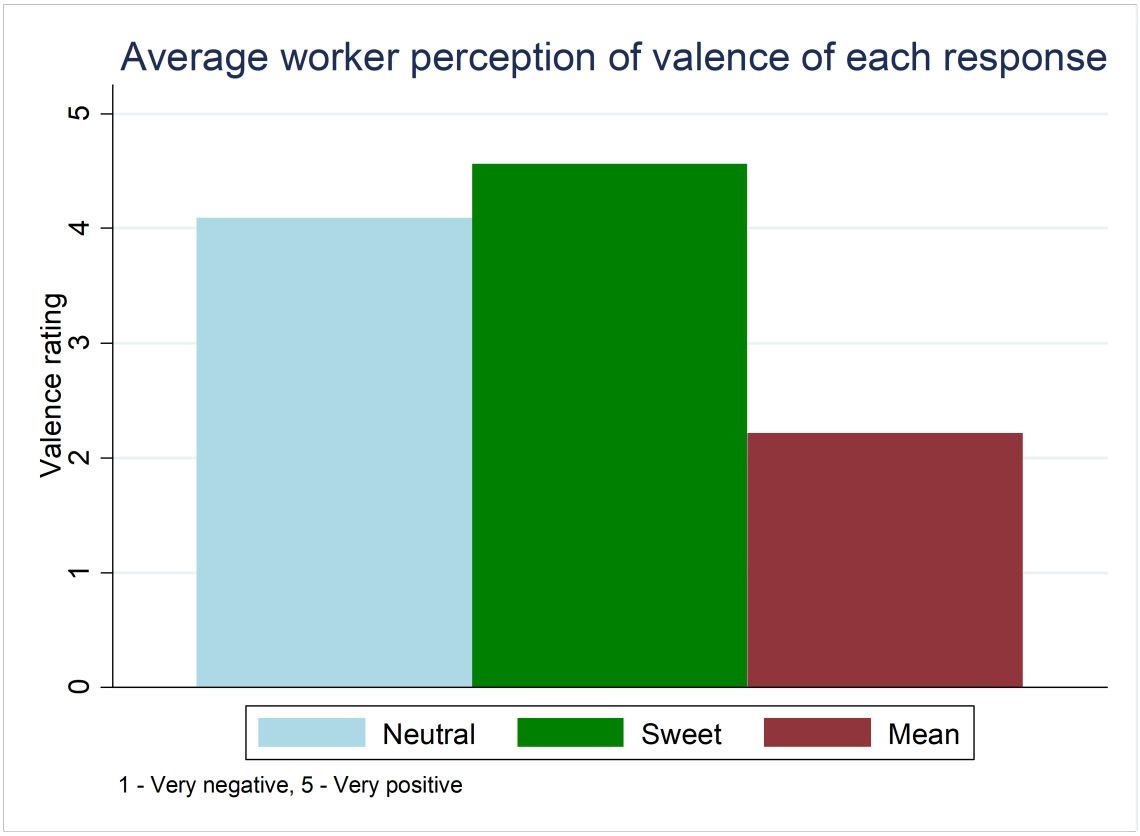

Figure B.2: Average worker perception of each response on a negative to positive 5-point Likert scale.

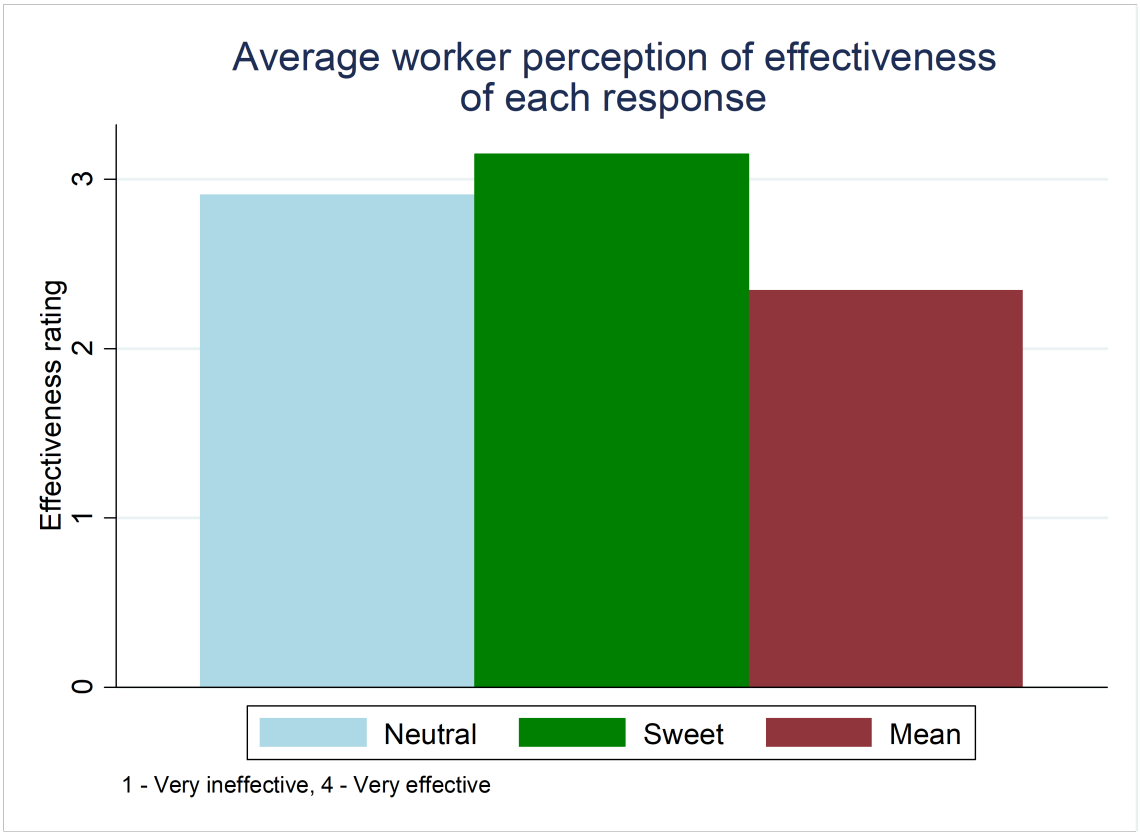

Figure B.3: Average worker perception of each response on an ineffective to effective 4-point Likert scale.

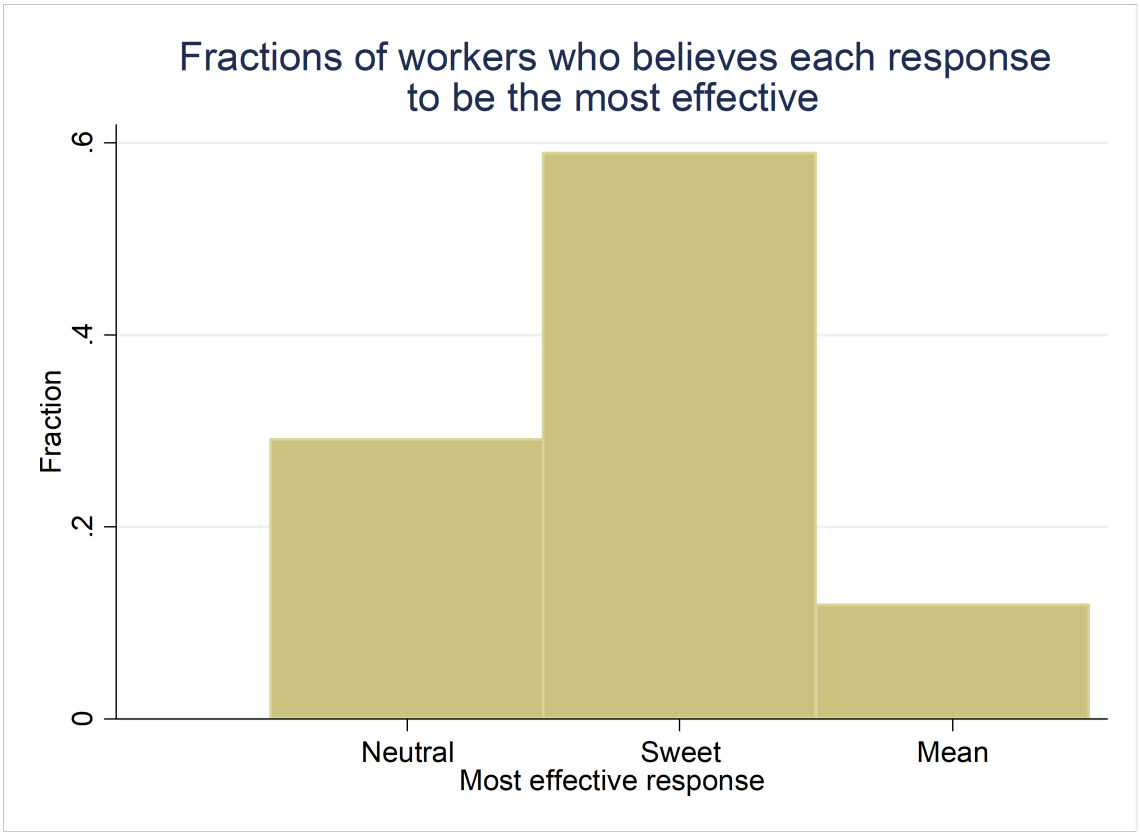

Figure B.4: Fractions of workers who rated each response to be the most effective response.

have any tied "most effective" rating (58.98% of the subjects). Figure 6 presents the histogram of the fraction of workers who find each of these responses to be the most effective. Workers appear to believe that the "sweet" response is the most effective, that the "neutral" response is the next effective, and the "mean" response is the least effective. These differences are all significant (tests of proportions, two-tailed, p-value < 0.01).

In this next analysis, I limit the data to worker As. Figure B.5 presents the histograms of the fractions of worker As that rated each response as the most effective. The difference in the proportions of workers' ratings across these conditions are still significant (tests of proportions, two-tailed, p-value < 0.01, with the exception of the "neutral" and "mean" response in the "A cheated more" disclosure condition, which

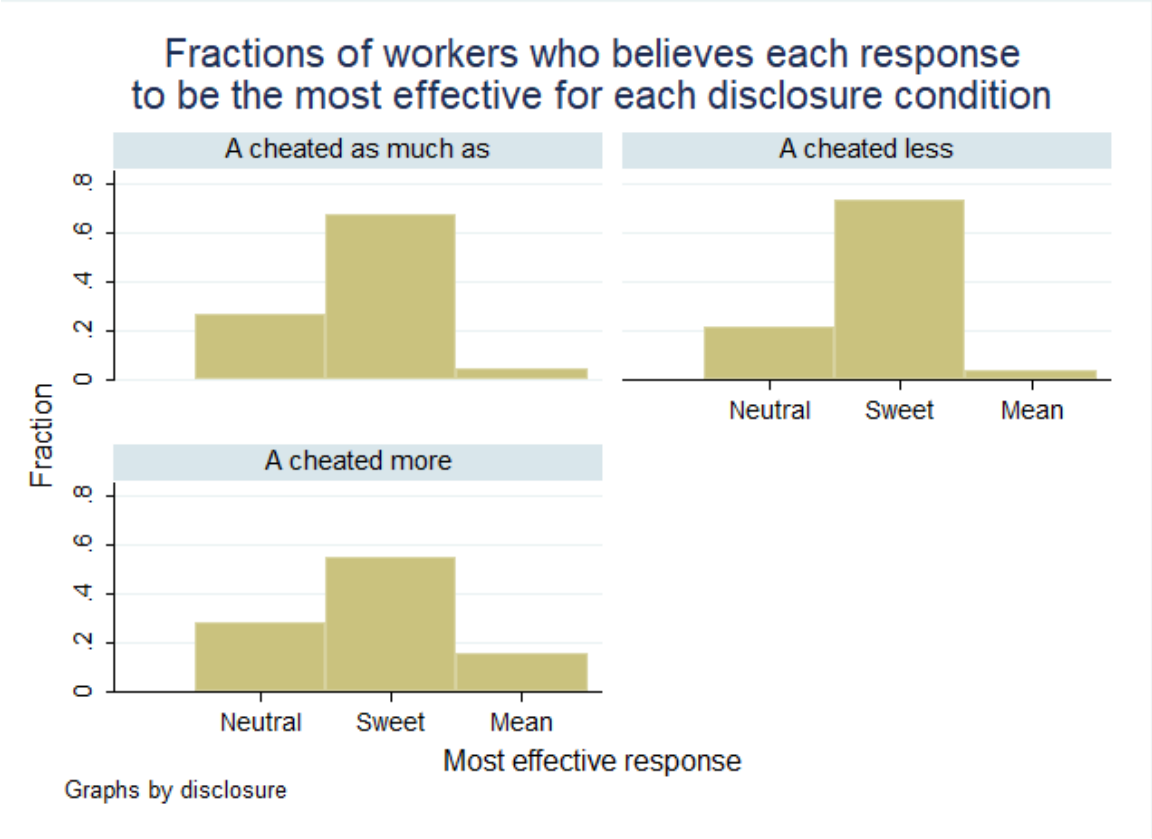

Figure B.5: Fractions of worker As who rated each response to be the most effective response by the performance disclosure condition that the worker is in.

is p-value < 0.05). The "sweet" response is still rated as the most effective response across all three performance disclosure conditions.

However, workers As' perceptions of a response's effectiveness depend on the preceding performance disclosure condition. While there are no significant differences in the proportions of workers who believe the "neutral" response is the most effective across the performance disclosure conditions, some differences exist for the other disclosure conditions. A larger proportion of workers rate the "sweet" response to be the most effective response in the "A cheated less" disclosure condition compared to the "A cheated as much as" disclosure condition (test of proportions, two-tailed, p-value < 0.10) as well as the "A cheated more" disclosure condition (test of proportions, two-tailed, p-value < 0.01). In addition, a larger proportion of workers rate that the "mean" response is the most effective response in the "A cheated more" disclosure condition relative to the "A cheated as much as" disclosure condition (test of proportions, two-tailed, p-value < 0.05) and to the "A cheated less" disclosure condition (test of proportions, two-tailed, p-value < 0.01).

B.4 Additional results of manager experiment post-performance disclosure

B.4.1 Comparing managers' post-information beliefs to Bayesian predictions

Because this elicitation procedure asked that the managers give a distribution of their beliefs of the number of times that A cheated, I can also compare the managers' actual expectations to the Bayesian predictions. Given the literature reviewed earlier, I hypothesize that

Hypothesis 1 (**Post-information: Comparison to Bayesian**). *Managers' posterior beliefs about the number of times that worker A cheated are consistent with what is predicted by Bayes' Rule.*

For each manager, I calculate his or her posteriors according to the Bayes' Rule, given his or her priors pre-performance disclosure. I then take the expected number of times that A cheated under this distribution and call this the "Bayesian posterior mean."

Figure B.6 plots the managers' actual posterior means on the Bayesian posterior means. The y-axis (the "manager's reported posterior mean") are managers' expectations of the number of times worker A cheated post-performance disclosure, a value between 0 and 5. The x-axis (the "Bayesian posterior mean") is what each of the manager's expectations ought to be if that manager applied Bayes' Rule, given his or her prior. The green dots and lines are the posterior means and the fitted line for managers in the "A cheated less" disclosure condition. Similarly, the red dots and lines are the posterior means and fitted line for managers in the "A cheated more" disclosure condition.

If managers are applying the Bayes' Rule exactly, then the dots ought to cluster along a 45-degree line (e.g. managers expecting worker A to have cheated exactly 1

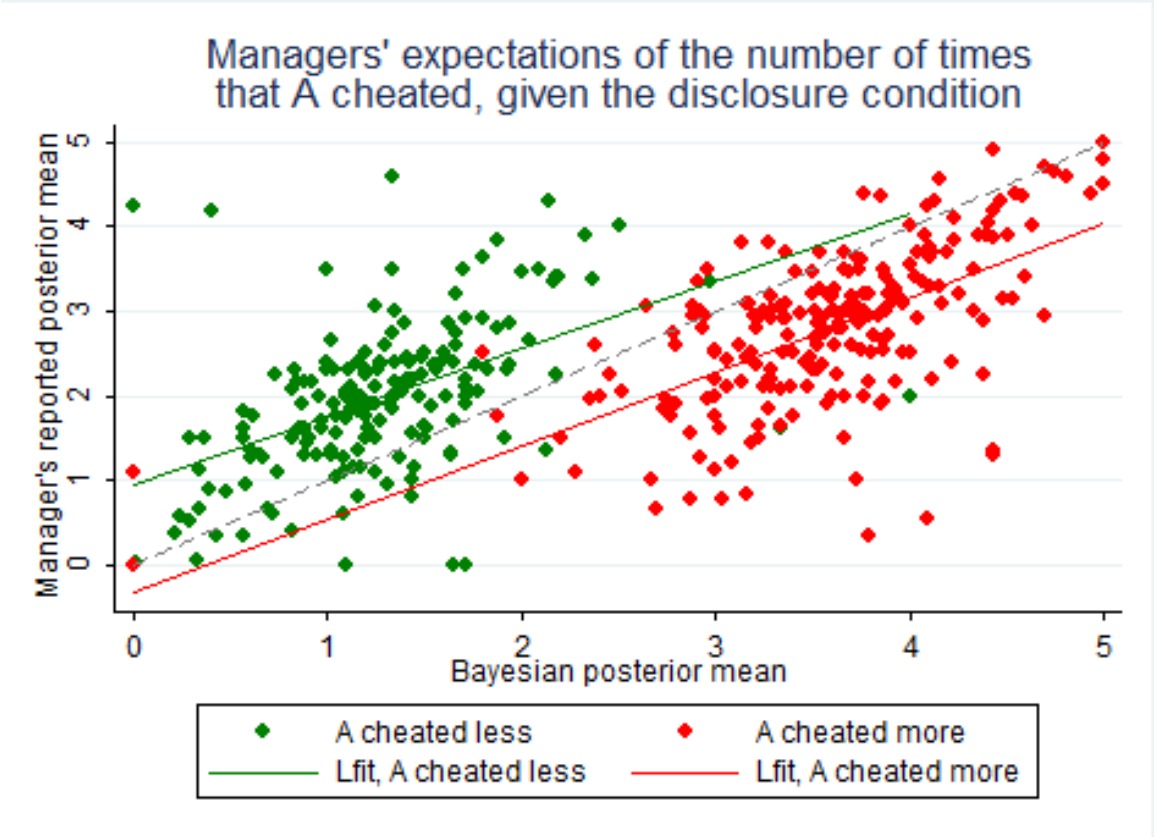

Figure B.6: Scatter plot of the managers' calculated Bayesian posterior means and their actual reported posterior means with fitted lines.

time are managers who, by Bayes' Rule, should expect that worker A cheated exactly 1 time).

Visually, managers do not appear to adjust their expectations closely to what Bayes' rule predicts given each manager's prior. I perform a series of t-test testing for differences between the manager's reported posterior mean and the Bayesian posterior mean. I find that managers in the "A cheated less" condition report posterior means that are under-adjusted to positive news relative to the Bayesian prediction (t-test, two-tailed: p-value < 0.01). In contrast, managers in the "A cheated more" condition report posterior means that are also under-adjusted to the "bad" news relative to the Bayesian prediction (t-test, two-tailed: p-value < 0.01). This systematic under-adjusting is consistent with existing studies (Edwards, 1968; Coutts, 2018) that find that people update like conservative Bayesians.

Result 1 (**Post-information: Comparison to Bayesian**). *Managers' posterior beliefs about the number of times that A cheated are inconsistent with the Bayesian benchmark. In addition, managers are under-adjusting post-performance disclosure, relative to the Bayesian prediction.*

B.4.2 Change in managers' post-performance disclosure beliefs

I consider the effect that good ("A cheated less") and bad ("A cheated more") performance disclosure has on the managers' beliefs. While there is evidence in asymmetric updating of beliefs when the "good" or "bad" information is ego-relevant, *Coutts* (2019) finds evidence of asymmetrical updating in beliefs even when the information is not ego-relevant.

Hypothesis 2 (**Post-information: Symmetric belief updates**). *Managers' beliefs about the number of times that worker A cheated update symmetrically across "A cheated less" versus "A cheated more" disclosures.*

Table B.3: OLS regression of the reported posterior means on the dummy variable for the performance disclosure condition, the Bayesian posterior mean, and the interaction term.

	Post-performance disclosure: Reported posterior means
A cheated less	1.283***
	(0.364)
Bayesian posterior mean	0.874***
	(0.082)
A cheated less X Bayesian posterior mean	-0.072
	(0.183)
Constant	-0.33
	(0.297)
Observations	414
R-squared	0.454

Robust standard errors in parentheses
*** p< 0.01, ** p< 0.05, * p< 0.1

The OLS regression (Table B.3) regresses the posterior means on a dummy variable for whether the manager is in the "A cheated less" condition or the "A cheated more" condition, the manager's Bayesian posterior mean, and their interaction. If managers update their beliefs asymmetrically across the two performance disclosure conditions, then the coefficient on the interaction ("A cheated less X Bayesian posterior mean") should be significant. The intuition for this is that for each point increase in the Bayesian posterior mean, a manager's reported posterior mean should increase/decrease at different rates depending on whether that manager is in the "A cheated less" condition or the "A cheated more" condition. Instead, the coefficient on this interaction term is small and insignificant.[1]

[1] *Eil and Rao* (2011) also note that people tend to respond more predictably when they see "good" news compared to "bad" news only when the information is ego-relevant, but not otherwise. In this context, it means that there should not be a difference in posterior means and variance across the two performance disclosure conditions. The variances of the managers' reported posterior means do not differ in the "A cheated less" and "A cheated more" performance disclosure conditions (variance ratio test, one-tailed: p-value = 0.3636). In addition, I perform a variance ratio test of the residuals of OLS regression fitting the Bayesian posterior predictions to the managers' reported posterior means. The variance ratio test of the residuals of the OLS regressions fitting the Bayesian posterior

Result 2 (**Post-information: Symmetric belief updates**). *Managers' posterior beliefs update symmetrically across "A cheated less" and "A cheated more" disclosure conditions.*

predictions to the managers' reported posterior means is also not significant (variance ratio test, one-tailed: p-value = 0.7627).

BIBLIOGRAPHY

BIBLIOGRAPHY

Abeler, J., J. Calaki, K. Andree, and C. Basek (2010), The power of apology, *Economics Letters*, *107*(2), 233–235.

Acquisti, A. (2004), Privacy in electronic commerce and the economics of immediate gratification, in *Proceedings of the 5th ACM Conference on Electronic Commerce*, EC '04, pp. 21–29, ACM, New York, NY, USA, doi:10.1145/988772.988777.

Acquisti, A., and R. Gross (2006), Imagined communities: Awareness, information sharing, and privacy on the facebook, in *Privacy Enhancing Technologies*, pp. 36–58, Springer.

Acquisti, A., and R. Gross (2009), Predicting social security numbers from public data, *Proceedings of the National Academy of Sciences*, *106*(27), 10,975–10,980.

Acquisti, A., and J. Grossklags (2005), Privacy and rationality in individual decision making, *IEEE Security & Privacy*, (1), 26–33.

Acquisti, A., L. K. John, and G. Loewenstein (2012), The impact of relative standards on the propensity to disclose, *Journal of Marketing Research*, *49*(2), 160–174.

Acquisti, A., L. Brandimarte, and G. Loewenstein (2015), Privacy and human behavior in the age of information, *Science*, *347*(6221), 509–514, doi:10.1126/science.aaa1465.

Adar, E., D. S. Tan, and J. Teevan (2013), Benevolent deception in human computer interaction, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pp. 1863–1872, ACM, New York, NY, USA, doi:10.1145/2470654.2466246.

Afridi, F., S. X. Li, and Y. Ren (2015), Social identity and inequality: The impact of china's *hukou* system, *Journal of Public Economics*, *123*, 17–29, doi:10.1016/j.jpubeco.2014.12.011.

Aho, K., D. Derryberry, and T. Peterson (2014), Model selection for ecologists: the worldviews of aic and bic, *Ecology*, *95*(3), 631–636.

Akerlof, G. A., and R. E. Kranton (2000), Economics and identity, *The Quarterly Journal of Economics*, *115*(3), 715–753, doi:10.1162/003355300554881.

Akerlof, G. A., and R. E. Kranton (2005), Identity and the economics of organizations, *Journal of Economic Perspectives*, *19*(1), 9–32, doi:10.1257/0895330053147930.

Allison, S. T., J. K. Beggan, and E. H. Midgley (1996), The quest for" similar instances" and" simultaneous possibilities": Metaphors in social dilemma research., *Journal of Personality and Social Psychology*, *71*(3), 479.

Andersson, O., M. M. Galizzi, T. Hoppe, S. Kranz, K. Van Der Wiel, and E. Wengström (2010), Persuasion in experimental ultimatum games, *Economics Letters*, *108*(1), 16–18.

Andreoni, J. (1995), Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments, *The Quarterly Journal of Economics*, *110*(1), 1–21.

Asch, S. E. (1956), Studies of independence and conformity: A minority of one against a unanimous majority, *Psychological Monographs: General and Applied*, *70*(9), 1–70.

Bandiera, O., I. Barankay, and I. Rasul (2005), Social preferences and the response to incentives: Evidence from personnel data, *The Quarterly Journal of Economics*, *120*(3), 917–962, doi:10.1093/qje/120.3.917.

Banerjee, R. (2016), On the interpretation of bribery in a laboratory corruption game: moral frames and social norms, *Experimental Economics*, *19*(1), 240–267, doi:10.1007/s10683-015-9436-1.

Bardsley, N. (2008), Dictator game giving: altruism or artefact?, *Experimental Economics*, *11*(2), 122–133, doi:10.1007/s10683-007-9172-2.

Bardsley, N., and R. Sausgruber (2005), Conformity and reciprocity in public good provision, *Journal of Economic Psychology*, *26*(5), 664–681.

Barr, A., T. Lane, and D. Nosenzo (2018), On the social inappropriateness of discrimination, *Journal of Public Economics*, *164*, 153–164.

Barrick, M. R., J. A. Shaffer, and S. W. DeGrassi (2009), What you see may not be what you get: relationships among self-presentation tactics and ratings of interview and job performance., *Journal of applied psychology*, *94*(6), 1394.

Benjamin, D. J., J. J. Choi, and A. J. Strickland (2010), Social identity and preferences, *American Economic Review*, *100*(4), 1913–1928, doi:10.1257/aer.100.4.1913.

Berlin, N., and M.-P. Dargnies (2016), Gender differences in reactions to feedback and willingness to compete, *Journal of Economic Behavior & Organization*, *130*, 320–336.

Bicchieri, C. (2005), *The grammar of society: The nature and dynamics of social norms*, Cambridge University Press, Cambridge.

Bicchieri, C., and A. Chavez (2010), Behaving as expected: Public information and fairness norms, *Journal of Behavioral Decision Making*, *23*(2), 161–178.

Blount, S., and R. P. Larrick (2000), Framing the game: Examining frame choice in bargaining, *Organizational Behavior and Human Decision Processes*, *81*(1), 43–71.

Blume, A., and A. Ortmann (2007), The effects of costless pre-play communication: Experimental evidence from games with pareto-ranked equilibria, *Journal of Economic theory*, *132*(1), 274–290.

Bonaccio, S., J. O'Reilly, S. L. O'Sullivan, and F. Chiocchio (2016), Nonverbal behavior and communication in the workplace: A review and an agenda for research, *Journal of Management*, *42*(5), 1044–1074.

Boskin, M. J. (1974), A conditional logit model of occupational choice, *Journal of Political Economy*, *82*(2), 389–398.

Bosman, R., and F. Van Winden (2002), Emotional hazard in a power-to-take experiment, *The Economic Journal*, *112*, 147–169, doi:10.1111/1468-0297.0j677.

boyd, d., and E. Hargittai (2010), Facebook privacy settings: Who cares?, *First Monday*, *15*(8).

Brandts, J., and C. Schwieren (2009), Frames and economic behavior: An experimental study, working paper.

Brewer, M. B., and R. M. Kramer (1986), Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing, *Journal of Personality and Social Psychology*, *50*(3), 543–549.

Brooks, D. J., and M. Murov (2012), Assessing accountability in a post-citizens united era: The effects of attack ad sponsorship by unknown independent groups, *American Politics Research*, *40*(3), 383–418.

Brunswik, E. (1955), Representative design and probabilistic theory in a functional psychology., *Psychological review*, *62*(3), 193.

Butler, J. V. (2014), Trust, truth, status and identity: An experimental inquiry, *The BE Journal of Theoretical Economics*, *14*(1), 293–338.

Cappelen, A. W., U. H.Nielsen, E. Ø. Sørensen, B. Tungodden, and J.-R. Tyran (2013), Give and take in dictator games, *Journal of Behavioral Decision Making*, *118*(2), 280–283, doi:10.1016/j.econlet.2012.10.030.

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

Charness, G. (2000), Self-serving cheap talk: A test of aumann's conjecture, *Games and Economic Behavior*, *33*(2), 177–194.

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

Charness, G., and B. Grosskopf (2004), What makes cheap talk effective? experimental evidence, *Economics Letters*, *83*(3), 383–389.

Charness, G., and M. Rabin (2002), Understanding social preferences with simple tests, *Quarterly Journal of Economics*, *117*(3), 817–869, doi:10.1162/003355302760193904.

Charness, G., L. Rigotti, and A. Rustichini (2007), Individual behavior and group membership, *American Economic Review*, *97*(4), 1340–1352, doi:10.1257/aer.97.4.1340.

Charness, G., R. Cobo-Reyes, and N. Jiménez (2014), Identities, selection, and contributions in a public-goods game, *Games and Economic Behavior*, *87*, 322–338.

Chen, R., and Y. Chen (2011), The potential of social identity for equilibrium selection, *American Economic Review*, *101*(6), 2562–2589, doi:10.1257/aer.101.6.2562.

Chen, Y., and S. X. Li (2009), Group identity and social preferences, *American Economic Review*, *99*(1), 431–457, doi:10.1257/aer.99.1.431.

Chen, Y., S. X. Li, T. X. Liu, and M. Shih (2014), Which hat to wear? impact of natural identities on coordination and cooperation, *Games and Economic Behavior*, *84*, 58–86, doi:10.1016/j.geb.2013.12.002.

Chowdhury, S. M., J. Y. Jeon, and B. Saha (2017), Gender differences in the giving and taking variants of the dictator game, *Southern Economic Journal*, *84*(2), 474–483, doi:10.1002/soej.12223.

Cialdini, R. B., R. R. Reno, and C. A. Kallgren (1990), A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places., *Journal of Personality and Social Psychology*, *58*(6), 1015.

Conti, G., and E. Sobiesk (2010), Malicious interface design: Exploiting the user, in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 271–280, ACM, New York, NY, USA, doi:10.1145/1772690.1772719.

Cookson, R. (2000), Framing effects in public goods experiments, *Experimental Economics*, *3*(1), 55–79, doi:10.1007/BF01669207.

Cooper, R., D. V. DeJong, R. Forsythe, and T. W. Ross (1992), Communication in coordination games, *The Quarterly Journal of Economics*, *107*(2), 739–771.

Coutts, A. (2019), Good news and bad news are still news: Experimental evidence on belief updating, *Experimental Economics*, *22*(2), 369–395.

Craig, S. C., P. S. Rippere, and M. S. Grayson (2014), Attack and response in political campaigns: An experimental study in two parts, *Political Communication*, *31*(4), 647–674.

Crawford, V. (1998), A survey of experiments on communication via cheap talk, *Journal of Economic theory*, *78*(2), 286–298.

Crumlish, C., and E. Malone (2009), *Designing social interfaces: Principles, patterns, and practices for improving the user experience*, " O'Reilly Media, Inc.".

D'Adda, G., M. Drouvelis, and D. Nosenzo (2015), Norm eliciation in within-subject designs: Testing for order effects, ceDEx Discussion Paper Series ISSN 1749-3293.

Dana, J., R. A. Weber, and J. X. Kuang (2007), Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness, *Economic Theory*, *33*(1), 67–80.

Davies, P. S., M. J. Greenwood, and H. Li (2001), A conditional logit approach to u.s. state-to-state migration, *Journal of Regional Science*, *41*(2), 337–360, doi: 10.1111/0022-4146.00220.

Debatin, B., J. P. Lovejoy, A.-K. Horn, and B. N. Hughes (2009), Facebook and online privacy: Attitudes, behaviors, and unintended consequences, *Journal of Computer-Mediated Communication*, *15*(1), 83–108, doi:10.1111/j.1083-6101.2009.01494.x.

Deshpande, S. P., P. P. Schoderbek, and J. Joseph (1994), Promotion decisions by managers: A dependency perspective, *Human Relations*, *47*(2), 223–232.

Deutsch, M., and H. B. Gerard (1955), A study of normative and informational social influences upon individual judgment., *The Journal of Abnormal and Social Psychology*, *51*(3), 629.

DeVaro, J. (2006), Internal promotion competitions in firms, *The Rand Journal of Economics*, *37*(3), 521–542.

Dowling, C. M., and A. Wichowsky (2015), Attacks without consequence? candidates, parties, groups, and the changing face of negative advertising, *American Journal of Political Science*, *59*(1), 19–36.

Dreber, A., T. Ellingsen, M. Johannesson, and D. G. Rand (2013), Do people care about social context? framing effects in dictator games, *Experimental Economics*, *16*(3), 349–371, doi:10.1007/s10683-012-9341-9.

Drogosz, L. M., and P. E. Levy (1996), Another look at the effects of appearance, gender, and job type on performance-based decisions, *Psychology of Women Quarterly*, *20*(3), 437–445.

Duffy, J., and N. Feltovich (2002), Do actions speak louder than words? an experimental comparison of observation and cheap talk, *Games and Economic Behavior*, *39*(1), 1–27.

Dufwenberg, M., S. Gächter, and H. Hennig-Schmidt (2011a), The framing of games and the psychology of play, *Games and Economic Behavior*, *73*(2), 459–478.

Dufwenberg, M., S. Gächter, and H. Henning-Schmidt (2011b), The framing of games and the psychology of strategic choice, *Games and Economic Behavior*, *73*(2), 459–478.

Eckel, C. C., and P. J. Grossman (2005), Managing diversity by creating team identity, *Journal of Economic Behavior & Organization*, *58*(3), 371–392, doi:10.1016/j.jebo.2004.01.003.

Eichenberger, R., and F. Oberholzer-Gee (1998), Rational moralists: The role of fairness in democratic economic politics, *Public Choice*, *94*(1-2), 191–210.

Eil, D., and J. M. Rao (2011), The good news-bad news effect: asymmetric processing of objective information about yourself, *American Economic Journal: Microeconomics*, *3*(2), 114–38.

Ellingsen, T., and E. Mohlin (2014), Situations and norms, working paper.

Ellingsen, T., M. Johannesson, J. Mollerstrom, and S. Munkhammar (2012), Social framing effects: Preferences or beliefs?, *Games and Economic Behavior*, *76*(1), 117–130, doi:10.1016/j.geb.2012.05.007.

Erkut, H., D. Nosenzo, and M. Sefton (2015), Identifying social norms using coordination games: Spectators vs. stakeholders, *Economics Letters*, *130*, 28–31.

Farrow, K., G. Grolleau, and L. Ibanez (2018), Designing more effective norm interventions: The role of valence, cEE-M working papers 18-15, CEE-M, Universitiy of Montpellier, CNRS, INRA, Montpellier SupAgro.

Fehr, E., and K. M. Schmidt (1999), The theory of fairness, competition, and cooperation, *Quarterly Journal of Economics*, *114*(3), 817–868, doi:10.1162/003355399556151.

Fischbacher, U., and V. Utikal (2013), On the acceptance of apologies, *Games and Economic Behavior*, *82*, 592–608.

Gächter, S., and E. Renner (2003), Leading by example in the presence of free rider incentives, *Tech. rep.*, University of St. Gallen.

Gächter, S., D. Nosenzo, and M. Sefton (2013), Peer effects in pro-social behavior: social norms or social preferences?, *Journal of the European Economic Association*, *11*(3), 548–573.

Gächter, S., L. Gerhards, and D. Nosenzo (2017), The importance of peers for compliance with norms of fair sharing, *European Economic Review*, *97*, 72–86.

Gallup, A. (1991), Gallup and party id: Birth of a question, *Public Perspective*, *2*(5), 23–24, interview.

Gangadharan, L., T. Jain, P. Maitra, and J. Vecci (2016), Social identity and governance: the behavioral response to female leaders, *European Economic Review*, *90*, 302–325.

Garramone, G. M. (1985), Effects of negative political advertising: The roles of sponsor and rebuttal, *Journal of Broadcasting & Electronic Media*, *29*(2), 147–159.

Gerber, A. S., and T. Rogers (2009), Descriptive social norms and motivation to vote: Everybody's voting and so should you, *The Journal of Politics*, *71*(1), 178–191.

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

Gneezy, A., U. Gneezy, G. Riener, and L. D. Nelson (2012), Pay-what-you-want, identity, and self-signaling in markets, *Proceedings of the National Academy of Sciences of the United States of America*, *109*(19), 7236–7240, doi:10.1073/pnas.1120893109.

Goerg, S. J., and G. Walkowitz (2010), On the prevalence of framing effects across subject-pools in a two-person cooperation game, *Journal of Economic Psychology*, *31*(6), 849–859, doi:10.1016/j.joep.2010.06.001.

Goerg, S. J., D. Rand, and G. Walkowitz (2017), Framing effects in the prisoner's dilemma but not in the dictator game, working.

Goette, L., D. Huffman, and S. Meier (2006), The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups, *American Economic Review*, *96*(2), 212–216, doi:10.1257/000282806777211658.

Goette, L., D. Huffman, and S. Meier (2012), The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups, *American Economic Journal: Microeconomics*, *4*(1), 101–115, doi:10.1257/mic.4.1.101.

Grossman, P., and C. Eckel (2012), Giving versus taking: A "real donation" comparison of warm glow and cold prickle, monash Economics working papers 40-12.

Grossman, P. J., and C. C. Eckel (2015), Giving versus taking for a cause, *Economics Letters*, *132*(1), 28–30, doi:10.1016/j.econlet.2015.04.002.

Halvorsen, T. U. (2015), Are dictators loss averse?, *Rationality and Society*, *27*(4), 469–491, doi:10.1177/1043463115605302.

Hardisty, D. J., E. J. Johnson, and E. U. Weber (2010), A dirty word or a dirty world? attribute framing, political affiliation, and query theory, *Psychological Science*, *21*(1), 86–92.

Harper, F. M., Yan Chen, J. Konstan, and S. X. Li (2010), Social comparisons and contributions to online communities: A field experiment on movielens, *The American economic review*, pp. 1358–1398.

Hauge, K. E., K. A. Brekke, L.-O. Johansson, O. Johansson-Stenman, and H. Svedsäter (2016), Keeping others in our mind or in our heart? distribution games under cognitive load, *Experimental Economics*, *19*(3), 562–576, doi: 10.1007/s10683-015-9454-z.

Higgins, C. A., T. A. Judge, and G. R. Ferris (2003), Influence tactics and work outcomes: A meta-analysis, *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *24*(1), 89–106.

Hitt, M. A., and S. H. Barr (1989), Managerial selection decision models: Examination of configural cue processing., *Journal of Applied Psychology*, *74*(1), 53.

Hoffman, S. D., and G. J. Duncan (1988), Multinomial and conditional logit discrete-choice models in demography, *Demography*, *25*(3), 415–427, doi:10.2307/2061541.

Horton, J. J., D. G. Rand, and R. J. Zeckhauser (2011), The online laboratory: Conducting experiments in a real labor market, *Experimental Economics*, *14*(3), 399–425.

Hosoda, M., E. F. Stone-Romero, and G. Coats (2003), The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies, *Personnel psychology*, *56*(2), 431–462.

Huff, C., and D. Tingley (2015), "who are these people?": Evaluating the demographic characteristics and political preferences of mturk survey respondents, *Research and Politics*, *2*(3), 1–12.

Iyengar, S., and S. J. Westwood (2015), Fear and loathing across party lines: New evidence on group polarization, *American Journal of Political Science*, *59*(3), 690–707, doi:10.1111/ajps.12152.

Kahneman, D. (2000), Preface, in *Choices, values, and frames*, edited by D. Kahneman and A. Tversky, pp. ix–xviii, Cambridge University Press, Cambridge.

Karni, E. (2009), A mechanism for eliciting probabilities, *Econometrica*, *77*(2), 603–606.

Kettner, S. E., and S. Ceccato (2014), Framing matters in gender-paired dictator games, working.

Kimbrough, E. O., and A. Vostroknutov (2016), Norms make preferences social, *Journal of the European Economic Association*, *14*(3), 608–638.

Knijnenburg, B. P., and A. Kobsa (2013), Helping users with information disclosure decisions: Potential for adaptation, in *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pp. 407–416, ACM, New York, NY, USA, doi:10.1145/2449396.2449448.

Knijnenburg, B. P., A. Kobsa, and H. Jin (2013), Counteracting the negative effect of form auto-completion on the privacy calculus.

Koch, J. W. (1998), Political rhetoric and political persuasion: the changing structure of citizens' preferences on health insurance during policy debate, *Public Opinion Quarterly*, *62*(2), 209–229.

Korenok, O., E. L. Millner, and L. Razzolini (2014), Taking, giving, and impure altruism in dictator games, *Experimental Economics*, *17*(3), 488–500.

Kranton, R., M. Pease, S. Sanders, and S. Huettel (2016), Groupy and non-groupy behavior: Deconstructing bias in social preferences, working paper, Duke University, Durham, NC.

Kranton, R. E., and S. G. Sanders (2017), Groupy versus non-groupy social preferences: personality, region, and political party, *American Economic Review*, *107*(5), 65–69.

Kraut, R. E., P. Resnick, and S. Kiesler (2012), *Evidence-based social design: Mining the social sciences to build online communities*, MIT Press.

Krupka, E., and R. A. Weber (2009), The focusing and informational effects of norms on pro-social behavior, *Journal of Economic Psychology*, *30*(3), 307–320.

Krupka, E., S. Leider, and M. Jiang (2015), A meeting of the minds: informal agreements and social norms, *Working paper*, University of Michigan.

Krupka, E. L., and R. A. Weber (2013a), Identifying social norms using coordination games: Why does dictator game sharing vary?, *Journal of the European Economic Association*, *11*(3), 495—524, doi:10.1111/jeea.12006.

Krupka, E. L., and R. A. Weber (2013b), Identifying social norms using coordination games: Why does dictator game sharing vary?, *Journal of the European Economic Association*, *11*(3), 495–524.

Krupka, E. L., S. Leider, and M. Jiang (2016), A meeting of the minds: Informal agreements and social norms, *Management Science*, *63*(6), 1708–1729.

Larrick, R. P., and S. Blount (1997), The claiming effect: Why players are more generous in social dilemmas than in ultimatum games., *Journal of Personality and Social Psychology*, *72*(4), 810.

Lau, R. R., and I. B. Rovner (2009), Negative campaigning, *Annual review of political science*, *12*, 285–306.

Lau, R. R., L. Sigelman, and I. B. Rovner (2007), The effects of negative political campaigns: a meta-analytic reassessment, *Journal of Politics*, *69*(4), 1176–1209.

Leibbrandt, A., P. Maitra, and A. Neelim (2015), On the redistribution of wealth in a developing country: Experimental evidence on stake and framing effects, *Journal of Economic Behavior & Organization*, *118*(1), 360–371, doi:10.1016/j.jebo.2015.02.015.

Leiserowitz, A., G. Feinberg, S. Rosenthal, N. Smith, A. Anderson, C. Roser-Renouf, and E. Maibach (2014), What's in a name? global warming vs. climate change, *Yale Project on Climate Change Communication, New Haven: CT*.

Li, S. X., K. Dogan, and E. Haruvy (2011), Group identity in markets, *International Journal of Industrial Organization*, *29*(1), 104–115, doi:10.1016/j.ijindorg.2010.04.001.

Liberman, V., S. M. Samuels, and L. Ross (2004), The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves, *Personality and social psychology bulletin*, *30*(9), 1175–1185.

List, J. A. (2007), On the interpretation of giving in dictator games, *Journal of Political Economy*, *115*(3), 482–493, doi:10.1086/519249.

London, M., and S. A. Stumpf (1983), Effects of candidate characteristics on management promotion decisions: An experimental study, *Personnel Psychology*, *36*(2), 241–259.

McCarter, M. W., and R. M. Sheremeta (2013), You can't put old wine in new bottles: The effect of newcomers on coordination in groups, *PLOS ONE*, *8*(1), doi:10.1371/journal.pone.0055058.

McCusker, C., and P. J. Carnevale (1995), Framing in resource dilemmas: Loss aversion and the moderating effects of sanctions, *Organizational Behavior and Human Decision Processes*, *61*(2), 190–201, doi:10.1006/obhd.1995.1015.

McFadden, D. (1974), Conditional logit analysis of qualitative choice behavior, in *Frontiers in Econometrics*, edited by P. Zarembka, pp. 105–142, Academic Press, New York, NY.

Mehta, J., C. Starmer, and R. Sugden (1994), The nature of salience: An experimental investigation of pure coordination games, *The American Economic Review*, *84*(3), 658–673.

Möbius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2014), Managing self-confidence.

Morrow, P. C., J. C. McElroy, B. G. Stamper, and M. A. Wilson (1990), The effects of physical attractiveness and other demographic characteristics on promotion decisions, *Journal of Management*, *16*(4), 723–736.

Nelson, T. E., R. A. Clawson, and Z. M. Oxley (1997a), Media framing of a civil liberties conflict and its effect on tolerance, *American Political Science Review*, *91*(3), 567–583, doi:10.2307/2952075.

Nelson, T. E., Z. M. Oxley, and R. A. Clawson (1997b), Toward a psychology of framing effects, *Political Behavior*, *19*(3), 221–246.

Nolan, J. M., P. W. Schultz, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius (2008), Normative social influence is underdetected, *Personality and social psychology bulletin*, *34*(7), 913–923.

Paolacci, G., J. Chandler, and P. G. Ipeirotis (2010), Running experiments on amazon mechanical turk, *Judgment and Decision Making*, *5*(5), 411–419.

Pathe, S. (2017), What's in a name? 'obamacare' vs. 'trumpcare' vs. 'ryancare', *Roll Call*, . `http://www.rollcall.com/news/politics/trumpcare-obamacare-ryancare`.

Pickup, M., E. O. Kimbrough, and E. A. de Rooji (2016), Experimental evidence on the role of identity and interest in voting behavior, working paper.

Pickup, M., E. O. Kimbrough, and E. A. de Rooji (2018a), Identity and the self-reinforcing effects of norm compliance, working paper.

Pickup, M., E. O. Kimbrough, and E. A. de Rooji (2018b), Expressive politics as (costly) norm following, working paper.

Rege, M., and K. Telle (2004), The impact of social approval and framing on cooperation in public good situations, *Journal of Public Economics*, *88*(7–8), 1625–1644, doi:10.1016/S0047-2727(03)00021-5.

Reinmuth, J. E., and M. D. Geurts (1975), The collection of sensitive information using a two-stage, randomized response model, *Journal of Marketing Research*, pp. 402–407.

Reinsch Jr, N. L., and J. A. Gardner (2014), Do communication abilities affect promotion decisions? some data from the c-suite, *Journal of Business and Technical communication*, *28*(1), 31–57.

Rosenbaum, S. M., S. Billinger, and N. Stieglitz (2014), Let's be honest: A review of experimental evidence of honesty and truth-telling, *Journal of Economic Psychology*, *45*, 181–196.

Roy, D. (1952), Quota restriction and goldbricking in a machine shop, *American Journal of Sociology*, *57*(5), 427–442.

Ruderman, M. N., and P. J. Ohlott (1994), *The realities of management promotion*, Center for Creative Leadership Greensboro, NC.

Ruderman, M. N., P. J. Ohlott, and K. E. Kram (1995), Promotion decisions as a diversity practice, *Journal of Management Development*, *14*(2), 6–23.

Ruderman, M. N., P. J. Ohlott, and K. E. Kram (1996), *Managerial promotion: The dynamics for men and women*, Center for Creative Leadership.

Rudman, L. A. (1998), Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management., *Journal of personality and social psychology*, *74*(3), 629.

Rugg, D. (1941), Experiments in wording questions: Ii, *Public Opinion Quarterly*, *5*(1), 91.

Schelling, T. C. (1980), *The strategy of conflict*, Harvard university press, Cambridge.

Schram, A., and G. Charness (2011), Social and moral norms in the laboratory, *UCSB manuscript*.

Schultz, P. W., J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius (2007), The constructive, destructive, and reconstructive power of social norms, *Psychological science*, *18*(5), 429–434.

Sell, J., and Y. Son (1997), Comparing public goods with common pool resources: Three experiments, *Social Psychology Quarterly*, *60*(2), 118–137.

Shang, J., and R. Croson (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, doi:10.1111/j.1468-0297.2009.02267.x.

Shih, M., T. L. Pittinsky, and N. Ambady (1999), Stereotype susceptibility: Identity salience and shifts in quantitative performance, *Psychological Science*, *10*(1), 81–84, doi:10.1111/1467-9280.00111.

Shih, M., T. L. Pittinsky, and A. Trahan (2006), Domain-specific effects of stereotypes on performance, *Self and Identity*, *5*(1), 1–14, doi:10.1080/15298860500338534.

Sonnemans, J., A. Schram, and T. Offerman (1998), Public good provision and public bad prevention: The effect of framing, *Journal of Economic Behavior & Organization*, *34*(1), 143–161, doi:10.1016/S0167-2681(97)00042-5.

Spiekermann, S., J. Grossklags, and B. Berendt (2001), E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior, in *Proceedings of the 3rd ACM Conference on Electronic Commerce*, EC '01, pp. 38–47, ACM, New York, NY, USA, doi:10.1145/501158.501163.

Steele, C. M., and J. Aronson (1995), Stereotype threat and the intellectual test performance of african americans, *Journal of Personality and Social Psychology*, *69*(5), 797–811, doi:10.1037/0022-3514.69.5.797.

Stevens, C. K., and A. L. Kristof (1995), Making the right impression: A field study of applicant impression management during job interviews., *Journal of applied psychology*, *80*(5), 587.

Stewart, G. L., S. L. Dustin, M. R. Barrick, and T. C. Darnold (2008), Exploring the handshake in employment interviews., *Journal of Applied Psychology*, *93*(5), 1139.

Stutzman, F., and J. Kramer-Duffield (2010), Friends only: Examining a privacy-enhancing behavior in facebook, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pp. 1553–1562, ACM, New York, NY, USA, doi:10.1145/1753326.1753559.

Stutzman, F., J. Vitak, N. Ellison, R. Gray, and C. Lampe (2012), Privacy in interaction: Exploring disclosure and social capital in facebook, in *Proceedings of the International AAAI Conference on Web and Social Media*, AAAI.

Sugden, R. (1995), A theory of focal points, *The Economic Journal*, *105*(430), 533–550.

Sukumaran, A., S. Vezich, M. McHugh, and C. Nass (2011), Normative influences on thoughtful online participation, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3401–3410, ACM.

Swope, K., J. Cadigan, P. Schmitt, and R. Shupp (2008), Social position and distributive justice: Experimental evidence, *Southern Economic Journal*, *74*(3), 811–818.

Tajfel, H., and J. C. Turner (1979), An integrative theory of intergroup conflict, in *The Social Psychology of Intergroup Relations*, edited by S. Worchel and W. G. Austin, pp. 33–47, Brooks/Cole, Monterey, CA.

Tedeschi, B. (2012), E-commerce report; everybody talks about online privacy, but few do anything about it.

Terry, D. J., and A. T. O'Brien (2001), Status, legitimacy, and ingroup bias in the context of an organizational merger, *Group Processes & Intergroup Relations*, *4*(3), 271–289, doi:10.1177/1368430201004003007.

Tews, M. J., K. Stafford, and J. Zhu (2009), Beauty revisited: The impact of attractiveness, ability, and personality in the assessment of employment suitability, *International Journal of Selection and Assessment*, *17*(1), 92–100.

Thaler, R. H., and C. R. Sunstein (2008), *Nudge*, Yale University Press.

Tourangeau, R., and T. Yan (2007), Sensitive questions in surveys., *Psychological bulletin*, *133*(5), 859.

Tsai, J. Y., S. Egelman, L. Cranor, and A. Acquisti (2011), The effect of online privacy information on purchasing behavior: An experimental study, *Information Systems Research*, *22*(2), 254–268.

Turkle, S. (1997), Multiple subjectivity and virtual community at the end of the freudian century, *Sociological inquiry*, *67*(1), 72–84.

Tversky, A., and D. Kahneman (1981), The framing of decisions and the psychology of choice, *Science*, *211*, 453–458.

van Dijk, E., and H. Wilke (2000), Decision-induced focusing in social dilemmas: Give-some, keep-some, take-some, and leave-some dilemmas, *Journal of Personality and Social Psychology*, *78*(1), 92–104, doi:10.1037/2F0022-3514.78.1.92.

Veselỳ, Š. (2015), Elicitation of normative and fairness judgments: Do incentives matter?, *Judgment and Decision Making*, *10*(2), 191–197.

Visser, M. S., and M. R. Roelofs (2011), Heterogenous preferences for altruism: gender and personality, social status, giving and taking, *Experimental Economics*, *14*(4), 490–506, doi:10.1007/s10683-011-9278-4.

Wang, Y., P. G. Leon, K. Scott, X. Chen, A. Acquisti, and L. F. Cranor (2013), Privacy nudges for social media: An exploratory facebook study, in *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pp. 763–770, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland.

Wang, Y., P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget, and N. Sadeh (2014), A field trial of privacy nudges for facebook, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pp. 2367–2376, ACM, New York, NY, USA, doi:10.1145/2556288.2557413.

Wash, R. (2010), Folk models of home computer security, in *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pp. 11:1–11:16, ACM, New York, NY, USA, doi:10.1145/1837110.1837125.

Weber, C., J. Dunaway, and T. Johnson (2012), It's all in the name: Source cue ambiguity and the persuasive appeal of campaign ads, *Political Behavior*, *34*(3), 561–584.

Weber, R. A. (2006), Managing growth to achieve efficient coordination in large groups, *American Economic Review*, *96*(1), 114–126, doi:10.1257/000282806776157588.

Wichardt, P. C. (2008), Identity and why we cooperate with those we do, *Journal of Economic Psychology*, *29*(2), 127–139, doi:10.1016/j.joep.2007.04.001.

Wiseman, D. B., and I. P. Levin (1996), Comparing risky decision making under conditions of real and hypothetical consequences, *Organizational Behavior and Human Decision Processes*, *66*(3), 241–250.

Young, S. D., and A. H. Jordan (2013), The influence of social networking photos on social norms and sexual health behaviors, *Cyberpsychology, Behavior, and Social Networking*, *16*(4), 243–247.

Zhang, L., and A. Ortmann (2012), On the interpretation of giving, taking, and destruction in dictator games and joy-of-destruction games, working paper.