Economic Inquiry

FOOLING MYSELF OR FOOLING OBSERVERS? AVOIDING SOCIAL PRESSURES BY MANIPULATING PERCEPTIONS OF DESERVINGNESS OF OTHERS

JAMES ANDREONI and ALISON SANCHEZ*

We present a novel experiment demonstrating strategies selfish individuals utilize to avoid social pressure to be altruistic. Subjects participate in a trust game, after which they have an opportunity to state their beliefs about their opponent's actions. Subsequently, subjects participate in a task designed to "reveal" their true beliefs. Subjects who initially made selfish choices falsely state their beliefs about their opponent's kindness. Their "revealed" beliefs were significantly more accurate, which exposed subjects' knowledge that their selfishness was unjustifiable by their opponent's behavior. The initial false statements complied with social norms, suggesting subjects' attempts to project a more favorable social image. (*JEL* C9, D03, D83)

I. INTRODUCTION

An increasingly common focus in social preference research is on the conditions which decrease individuals' willingness to act prosocially. Tension exists between the need for individuals to cooperate in social situations and their willingness to do so. In the past 10 years it has become increasingly clear that a significant, if not dominant, reason that individuals behave generously toward others is to maintain an image as a generous type of person (Andreoni and Bernheim 2009; Ariely, Bracha, and Meier 2009; Bénabou and Tirole 2006; Harbaugh 1998), or similarly, to avoid the guilt associated with not being generous (Battigalli and Dufwenberg 2013). Since signaling one's altruism often involves pooling among a range of both high and low altruism types, the cost of maintaining a good social or self-image can become rather expensive, especially for the least altruistic among those in the pool. Given this, it is natural to expect that if people anticipate a situation that may require a social signal of generosity, some may take opportunities to avoid having to provide such a signal, such as sidestepping someone asking for a charitable donation, especially when social image concerns are heightened (Andreoni, Rao, and Trachtman 2017; Dana, Cain, and Dawes 2006; Dana, Weber, and Kuang 2007; DellaVigna, List, and Malmendier 2012). Another way to see this phenomenon is that, if the person is unable to avoid the situation, sending a costly signal is better than sending no signal—a person will give to a charity if asked. But if a potential donor is given a choice, the person could prefer to not enter a game that will require a social signal of altruism—avoiding being asked in the first place yields even higher utility.

In this paper, we explore a different avenue by which people can avoid the expense of a social signal and limit the damage to their self- or social images. Rather than hiding from others by avoiding a situation that may require them to act altruistically in order to send a signal, they instead enter the situation, act selfishly, and subsequently manipulate the *perception* of their selfish action in order to signal their altruism. If one can send

*We thank Nageeb Ali, Richard Carson, Vincent Crawford, Matthew Goldman, Michael Kuhn, Mark Machina, Erin Troland, and seminar participants at the 2014 ESA Meetings, UC San Diego Economics Department, and UC San Diego Neural Interaction Lab for helpful discussions and comments. Paul Feldman and Vincent Leah-Martin provided excellent assistance in running the experiments. Financial support from the National Science Foundation (GRFP#2009081811) is gratefully acknowledged.

Andreoni: Professor of Economics, Department of Economics, University of California, San Diego, San Diego, CA 92093. Phone (858) 534-3832, Fax (858) 534-7040, E-mail andreoni@ucsd.edu

Sanchez: Assistant Professor of Economics, Department of Economics, University of San Diego School of Business, San Diego, CA 92110. Phone (619)260-4851, Fax (619) 260-4891, E-mail alisonsanchez@sandiego.edu

ABBREVIATION

QSR: Quadratic Scoring Rule

Economic Inquiry
(ISSN 0095-2583)
Vol. 58, No. 1, January 2020, 12–33

doi:10.1111/ecin.12777
Online Early publication April 13, 2019

a signal about their altruism by manipulating the perception of their selfishness then they are no longer bound to use their action as a signal of the altruism. In particular, suppose there is a possibility that the potential recipient of generosity is, objectively, either deserving of help or not. For example, donors can have different priors or different information that causes them to have different beliefs about the deservingness of charity recipients. If one can also credibly convey a belief that the potential recipient is undeserving, then this can weaken the signal conveyed by not being generous. So, rather than physically avoiding having to signal, the person can strategically avoid the need to signal by manipulating the beliefs held by others (and perhaps themselves) about the deservingness of the recipient.[1]

We explore this process of "perception manipulation" in a laboratory experiment. The study is designed to distinguish between the beliefs expressed purely for image concerns ("stated beliefs") and those indirectly expressed for pure monetary purposes ("revealed beliefs"). First, subjects make decisions in a modified trust game. Subsequently, we directly ask subjects to state the beliefs they held when those decisions were made. These "stated beliefs" are reported after their choices have been made, so the only reason to misrepresent them is to justify their choices to themselves or others. Thus, "stated" beliefs are the beliefs subjects express when *only* their image is at stake, but not their payoff. To determine whether and how the "stated" beliefs may differ from subjects' true underlying beliefs, we later ask subjects to make bets on possible game outcomes where their entire payoff depends on those bets. Concealed in this larger exercise is our ability to recover the beliefs that would justify the set of bets they chose. We term these "revealed beliefs." Since, as we anticipated, subjects did not recognize our intention to measure their beliefs from the "revealed belief" stage of the experiments, we can interpret these "revealed beliefs" as the true beliefs of the subjects.[2] The difference between their "stated" and "revealed" beliefs is our measure of belief manipulation.

1. Rabin (1993), Blount (1995), Fehr and Gächter (2000), Charness and Rabin (2002), Falk and Fischbacher (2006), and others have shown that people find it more acceptable to withhold generosity from those perceived to be selfish.

2. Informal postexperiment debriefings indicated that subjects were unaware that the revealed beliefs stage was in fact measuring their beliefs. See Section VI for a further justification of this assumption.

Our findings support the notion that belief manipulation is a significant and commonly employed strategy to avoid unfavorable image consequences of one's selfish actions. In particular, comparing the "stated" and "revealed" beliefs, we found that subjects who took more selfish actions systematically manipulated their "stated" beliefs to be more pessimistic about the likelihood that their partner in the game was altruistic and thus less deserving of kindness. By contrast, subjects who took more unselfish actions tended not to distort "stated" beliefs away from "revealed" beliefs. The unselfish subjects have no need to justify a selfish action and thus do not need to misreport their stated beliefs.

An important related paper by Di Tella et al. (2015) finds that players who acted selfishly toward their opponent were more likely to report that they believed that their opponent was selfish. The authors interpret this as evidence that people who wish to behave selfishly need to form negative beliefs about their "victims" in order to avoid aversive feelings of guilt associated with taking advantage of those who have been kind to them. Our study complements and improves upon theirs in two important ways. First, by measuring both revealed (true) beliefs and stated beliefs, we can measure the degree to which beliefs are being manipulated. Second, our study also opens up the possibility that selfish individuals, in addition to self-manipulation, are manipulating the beliefs of observers. In our study, it could be that subjects who act selfishly are using their stated beliefs as justification of their actions to themselves, and so have convinced themselves that the probability of meeting a trustworthy opponent is actually as small as they have actually stated. If self-signaling is a factor, as in Di Tella et al. (2015), we would expect there to be no difference between stated and revealed beliefs. If it is the case that selfish subjects need to "convince" themselves of their partner's selfishness then we expect to find that these players, having convinced themselves of their stated beliefs, will express the same self-serving beliefs when their payoff is at stake during the revealed beliefs elicitation. However, if social signaling is the dominant factor then we expect to see a difference between stated and revealed beliefs. In particular, we expect to find that selfish subjects purposefully underestimate the degree of kindness of their partner on their stated beliefs, but then when their payoff is at stake they will express their true "revealed" beliefs that their partner is kind.

We find that the selfish players have the largest difference between their stated and revealed beliefs, that this difference is self-serving in that selfish players systematically underestimate the degree of kindness of their partner, and that their revealed beliefs are significantly more accurate predictors of the actual behavior of their partners than their stated beliefs. This suggests to us that social rather than self-signaling is the main reason for perception manipulation.

The paper proceeds as follows: Section II provides background and motivation for the main hypotheses; Section III describes the experimental design; Section IV describes predictions; Section V presents the results; Section VI provides discussion; and Section VII concludes.

II. BACKGROUND AND MAIN HYPOTHESES

This study builds on growing literature on image motivation, social norms, and strategic avoidance of other-regarding behavior. Much of the previous research relies on an implicit assumption that individuals' only recourse in the face of the pressures associated with maintaining a self- or social image is to take an altruistic action or avoid the situation altogether. Three points of interest have emerged from previous studies that suggest alternative hypotheses.

First, individuals take into account how others perceive them when deciding to act prosocially. Charness and Dufwenberg (2006) elicited beliefs in one-shot public goods games and found that when players believe they are expected to be other regarding, they give according to what they believe others expect of them. This suggests that individuals can anticipate and are pressured by others' expectations of them. Another well-known finding is that compliance with social norms results in a positive social and self-image, whereas violating social norms results in a negative social and self-image (Akerlof and Kranton 2000; Benabou and Tirole 2002). Additionally, the consequences of violating group norms can loom large for some. A damaged social image may result in material punishment and in psychological costs to self-image.

Second, applying social pressure to comply with social norms has observable effects on behavior, but perhaps not an effect on the underlying preferences of an individual. Malmendier, te Velde, and Weber (2014) show that external motivators are a significant driver of sharing behavior. DellaVigna, List, and Malmendier (2012) find that the pressure that arises from publicly violating social norms is a driving force in a large number of charitable donations. Importantly, the authors also find that utility losses are significant, as about half of the donors in their study would have preferred to not donate, or to donate less. As such, a growing literature has documented that individuals actively seek strategies that can help them avoid opportunities to be other regarding. For example, a field experiment by Andreoni, Rao, and Trachtman (2017) finds that individuals physically avoid situations where they will be asked to give to charity. Both of these results confirm similar behaviors from lab studies (Broberg, Ellingsen, and Johannesson 2007; Dana, Cain, and Dawes 2006; Lazear, Malmendier, and Weber 2012).

The third pillar of evidence comes from Andreoni and Bernheim (2009). They experimentally manipulate the strength of a social signal of selfish behavior by randomly forcing some subjects to be perfectly selfish. They found that as the probability of being forced to be selfish went up, the incidence of voluntarily choosing to be perfectly selfish went up as well. This is direct evidence of a concern for social image apart from self-image as the main motivator of seemingly prosocial behavior.

In contrast to this, Rabin (1995) presented a model of self-deception in which individuals form self-serving beliefs that their own selfish actions are justified or otherwise not harmful to others. This self-deception then frees them to take a selfish action. In a study closely related to ours, Di Tella et al. (2015) find that people avoid altruistic actions by distorting beliefs about altruism. The authors conducted a modified dictator game where recipients had the opportunity to take a side payment in exchange for reducing the overall size of the pie. Dictators in this setting reported that recipients were likely selfish and used this self-serving belief to convince themselves to take selfish action against the recipients.

Comparing the studies of Andreoni and Bernheim on social image, and Di Tella et al. on self-image, it becomes immediately clear that self- and social image are linked. Andreoni and Bernheim's subjects may have had self-image concerns that then entered into their beliefs about how deserving the recipients may have been. Their experimental manipulations, however, work only on social image. Likewise, Di Tella et al.'s subjects have, at the very least, the experimenter as an audience and so subjects' behavior could to some degree also be capturing concerns for social image as well as self-image.

The manipulation could, therefore, contain elements of both. To fully understand the degree to which individuals are engaging in self- or social image manipulation, one would need to observe a measure of the manipulated belief about the behavior (and so deservingness) of their partners, their true belief, and then to compare these to the actual behavior. This leads to our main testable proposition of the paper:

A. *Main Hypothesis*

If manipulated beliefs (stated beliefs), true beliefs (revealed beliefs), and actual behavior of their partner are all in agreement, it suggests no belief manipulation. If both manipulated (stated) and true (revealed) beliefs are aligned, but both deviate systematically from actual behavior of their partner, it suggests self-image motivation is dominating social image motivation, that is, the subject both promotes and holds the same incorrect beliefs. However, if the manipulated beliefs (stated) differ systematically from the true beliefs (revealed), and the true beliefs (revealed) are most closely aligned with actual behavior of their partner, it points to social image motivation dominating self-image motivations. In particular, the subject is promoting beliefs known to be incorrect, but which, if believed by others, would strengthen their own social image.

Next we present a laboratory experiment designed to allow us to test this main hypothesis.

III. EXPERIMENTAL DESIGN

Each session consisted of three decision stages followed by one final payoff stage. In Stage 1, subjects play a modified Trust game with binary choices. We employ the strategy method: subjects were asked to make binding choices for different scenarios, and were instructed they would be paid based on one randomly chosen scenario at the end of the session. All choices were made with paper and pencil. In Stage 2, we collected nonincentivized stated beliefs. It is important that these elicited beliefs are nonincentivized. First, while there are no reasons to report honestly, there are plenty of reasons to report inaccurately for those who make selfish choices in Stage 1. If the failure to incentivize reports here results in inaccuracy, then one is hard-pressed to think of reasons other than belief manipulations that should systematically bias reports in one particular direction. In Stage 3, we collected incentivized revealed beliefs in a moderately complex manner to be explained. Players were paid in the last stage, Stage 4. As we describe the details we will refer to the diagram of the game in Figure 1. See Appendix B for the subject forms.

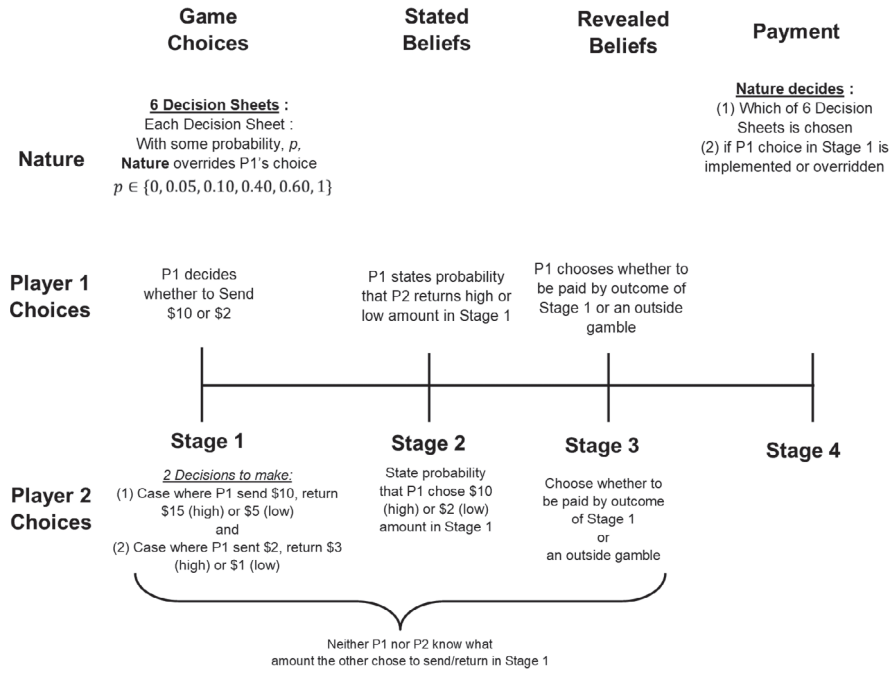
A. *Stage 1: Trust Game Choices*

We use a variation of the trust game of Berg, Dickhaut, and McCabe (1995) that restricts both players to two strategies. This facilitates the belief elicitation in Stage 3. In Stage 1 subjects were randomly divided into pairs and then randomly assigned to roles as Player 1 (P1) or Player 2 (P2). To begin, \$10 was placed into Player 1's "account." Player 1 then decided either to send the whole \$10 to Player 2 or to send \$2 to Player 2 and keep \$8 for themselves. Whichever amount that Player 1 chose to send to Player 2 was tripled by the experimenter. Player 2 decided how much of the tripled transfer they received, x, to return to Player 1. Player 2 decided between (1) whether to return $x/2$ to Player 1 and keep $x/2$, or (2) return $x/6$ to Player 1 and keep $5x/6$. Further, with probability $1-p$ Player 1's choice determined the amount transferred to Player 2 (either \$10 or \$2) and with probability p nature intervened and the Experimenter forced Player 1 to send the whole \$10 to Player 2. We examine choices for six different values of p, $p \in (0, .05, .20, .40, .60,$ and $1)$, with each player making choices on six corresponding "decision sheets." While the parameter p is common knowledge, there is no way that Player 2 could know or observe whether nature intervened. We elicited choices for all six values of p. Player 1 subjects made a total of six choices: one choice on each of six sheets by marking whether they would choose to send \$10 or \$2 to Player 2 for each sheet (even the treatment where $p=1$). By comparison, Player 2 subjects made 12 choices total: a conditional choice for the possibility that \$10 was sent, and, a conditional choice for the possibility that \$2 was sent for each value of p. When all decisions were completed by all subjects their decisions sheets were collected.

B. *Stage 2: Stated Beliefs*

While subjects knew at the beginning of the experiment that there would be three stages of the game, subjects did not know the exact nature of each stage until the beginning of each stage itself when directions for that particular stage only were given. Thus, all players made their decisions in Stage 1 before instructions were given for the remaining two stages of the game.

FIGURE 1
Timeline of Experiment



Notes: All decisions/choices are made simultaneously without knowledge of what the other player has chosen. Only after receiving payment in Stage 4 can players infer their opponents' moves.

In Stage 2 subjects were told, "We would like to know what you think the other player sent you." Each subject wrote their predictions of what the other player sent them on their own form called the Prediction Sheet, by writing in a percentage probability. It was made clear to subjects that there was no penalty or reward for accuracy and that their accuracy (or lack thereof) would not affect their final payoff.

In Stage 2, Prediction Sheets (Stated belief elicitation) were completed by both Player 1 and Player 2 subjects as follows.

Player 1 subjects were asked to predict the chances that Player 2 would return different amounts of money under different scenarios. It was publicly stated that if $10 was sent to Player 2, Player 2 could choose to send back either $15 or $5 to Player 1. If $2 was sent to Player 2 then Player 2 could return either $3 or $1 to Player 1. It was publicly known that Player 2 would be making a conditional choice for each possibility and that in each condition there existed a chance that nature could "override" Player 1's choice and force Player 1 to send the entire $10 to Player 2. Therefore, there were two attributions Player 2 could have made about who was responsible for sending the $10: either Player 1 was themselves responsible for sending the $10, or Player 1 was forced by the experimenter to send the $10. This stage was designed to test whether individuals operate on the assumption that their intentions will be taken into account when being judged by others. Therefore, Player 1 was asked to predict the chances that Player 2 sent back $15 or $5 for each of two possibilities: first, if Player 2 believed Player 1 was responsible for sending the $10 and second, if Player 2 believed that the experimenter was responsible for sending the $10 (i.e., Player 1 was forced by the experimenter). Player 1 was then asked to predict the chances that Player 2 sent $3 or $1 back to Player 1 under the $2 possibility. Player 1 made predictions for all six values of p. Again, it was made clear to subjects that there was no penalty or reward given for accuracy and that their statements would have no effect on their final payoff.

Player 2 subjects were simultaneously asked to predict the chances that Player 1 would send

them either \$10 or \$2. Player 2 knew that there were two separate ways to receive \$10: either when Player 1 decided to send the \$10, or when the experimenter forced Player 1 to send the \$10. We specifically addressed this issue in our instructions to Player 2. We made it clear to Player 2 that they should only predict the probability that Player 1 themselves chose to send the \$10 on Player 1's decision sheet, regardless of whether the experimenter later intervened. Thus, we instructed Player 2 to predict whether Player 1 marked \$10 on Player 1's decision sheet (and not the probability that Player 2 would *receive* \$10 from the experimenter). Player 2 made predictions for all six values of p.

C. Stage 3: Revealed Belief Elicitation

In Stage 3, we measured first-order beliefs about the action each player believed their opponent had taken. We used a modified multiple price list style approach to measure Player 1 and Player 2's preference between two payment options.[3] Subjects were informed that they would be making a series of decisions on "how they would like to be paid." On their "Payment Option Form," subjects chose between two payment options. Option 1 was termed the "Outcome of the Game." If players chose Option 1, they were paid based on the outcome of the trust game they played with their opponent. Subjects knew that the payment they would receive under this option depended in part on what amount the other player chose to send them, either $\$x/6$ (the "low" amount) or $\$5x/6$ (the "high" amount). By contrast, the second option, Option 2, did not depend on what the other player had sent them. Option 2 was purely an outside gamble with a q chance of receiving $\$x/6$ (the "low" amount) and a $1-q$ chance of receiving $\$5x/6$ (the "high" amount). Option 2 varied in incremental steps of 5%, which ranged from a 0% chance of $\$x/6$ and a 100% chance of $\$5x/6$, to a 100% chance of $\$x/6$ and a 0% chance of $\$5x/6$. Both options gave each player a chance of winning the same two amounts: either $\$x/6$ (the "low" amount) or $\$5x/6$ (the "high" amount). The sole difference between the two options was that the probability of receiving the high amount under Option 1 depended on the other player's action whereas under Option 2 the payment depended solely upon the chances subjects saw listed under Option 2. Therefore, the row at which a subject decides to switch from Option 2 (outside gamble) to Option 1 (Outcome of the Game) reveals the range of values of their belief about what the other player has chosen to send them. Subjects then filled out one payment option form for each of their six decision sheets.[4] Since the first row under Option 2 gives the player a 100% chance of receiving the high amount, subjects who understand the game structure should initially prefer Option 2, if they believe that there is less than a 100% chance they will receive the high amount under Option 1.[5]

It is common practice in experimental economics to elicit incentivized beliefs with a proper scoring rule, that is, one that is incentivized for truth telling. A popular technique is to use a quadratic scoring rule (QSR).[6] When subjects have private incentives to mislead us (or themselves) on their true beliefs, the QSR or any other devise that asks directly for beliefs can be expected to elicit biased reports from subjects, even if it is a proper scoring rule. We instead must derive true beliefs by masking them in another task which, without the subject's awareness, will indirectly reveal beliefs. Our revealed beliefs

3. Similar approaches can be found in Schlag, Tremewan, and Van der Weele (2015) and Schotter and Trevino (2014).

4. Previous price list style experiments have documented that a portion of subjects tend to switch multiple times between the two options presented (Holt and Laury 2002; Jacobson and Petrie 2009; Meier and Sprenger 2010). It is generally accepted that since multiple switch points can indicate subject confusion and are difficult to rationalize, a framing device may be necessary to avoid subject confusion and clarify the decision process (Andreoni and Sprenger 2012). We used animated instructions in order to illustrate the directions for the subjects. Out of 82 subjects, two subjects had multiple switch points on one or more of their payment option forms and one subject who switched "backwards" (starting with Option 1 and later switching to Option 2).

5. Under Option 2, the probability of receiving the high amount declines with each descending row, while the probability of receiving the low amount increases with each descending row. At the row where a subject believes that they would have a higher probability of receiving the high amount from the other player than the probability they see under Option 2, the subject has the incentive to switch to Option 1. Thus, the row where each subject switches allows us to infer their belief about the chances of the other player sending the high amount. In addition, we verbally instructed subjects that "Most people begin by preferring Option 2 and then switch to Option 1. Thus one way to view this task is to determine the best row to stop checking the box under Option 2 and start checking the box for Option 1."

6. Although the actual rule is often too difficult for subjects to be able to derive that truth telling is optimal, when simply instructed that truth telling is the optimal choice, it seems that subjects tend to trust the experimenter. Also see Andreoni and Mylovanov (2012) for a simpler and more direct implementation of a QSR in which subjects are incentivized by giving them the first derivative of the QSR and paying them the integral (sum of) bets up to the optimal odds.

task just described does exactly that. Being paid according to the outcome of the game is a risky gamble, and our instructions in Stage 3 of the study encourage subjects to view it in this way, focusing solely on expected payoffs from Stage 1 as compared to an abstract gamble in Stage 3. Given that they do not see the link to true beliefs, our technique for eliciting beliefs will be a proper scoring rule.[7] If by chance they do see the connection to revealed beliefs, their reports will be biased toward their stated beliefs, which will undermine our hypothesis. Informal post-experimental debriefings confirmed that subjects did not see through our guise. As we will report, we feel confident in our ability to separate and truthfully identify true (revealed) beliefs from stated beliefs.[8]

After collecting the Stage 3 decision forms, each subject was given a six-sided die. All of the Player 1s rolled dice to determine whether Player 1's decision would be implemented or overridden. In order to maintain anonymity as to which subjects were Player 1 and Player 2, all subjects rolled a die and recorded the number. The experimenter then rolled a die to determine which number on the die face would trigger a "forced choice" by the experimenter. Subjects' payments were put in a sealed envelope and handed to subjects as they left the experiment.

IV. PREDICTIONS

Following previous literature examining prosocial avoidance and excuses in second movers, we focus our analysis mainly on Player 2 behavior.[9] We propose that socially strategic selfish individuals deem that their selfish action will be evaluated in a more forgiving light if they are perceived as reacting to a belief that their opponent was selfish first, rather than if they are perceived as believing their opponent acted kindly toward them.[10] Since selfish individuals can no longer use their actions to signal their type to the experimenter, they must rely on the only means left available to maintain their social image: others' perceptions of their beliefs. Consequently, selfish individuals wishing to maintain their social image will indicate a stated belief that there is a low probability their opponent voluntarily sent the high amount and a high probability that their opponent sent them the low amount. This serves as an excuse for a selfish individual's behavior. However, we predict that not all of these individuals truly believe that their opponents were selfish. When it comes to receiving their final payoff, we posit that selfish individuals will be willing to risk their entire payment for the experiment on their true belief that their opponent was kind to them by sending the high amount. Thus, this would reveal that they believe they have a better chance of receiving the highest payoff from their opponent rather than from the outside gamble.[11] The intention to deceive others and manipulate perceptions requires strategic sophistication (Lisofsky et al. 2014). Not only must subjects understand the complexities of the underlying game structure, but also have the ability to anticipate others' beliefs and expectations about their behavior. Our experimental design has a built-in measure that allows us to identify subjects who understand the underlying game structure and are thus more strategically minded decision makers. We refer to these subjects as *sophisticated*.[12] We predict that if subjects have the higher-order reasoning skills necessary to execute the optimal choices in the revealed belief elicitation stage, then they will also be the subjects who possess the skills

7. We are confident that the framing of the stated belief task made salient that we were inquiring about their own beliefs. In contrast, the framing of the revealed beliefs task ("payment option form") made sure that the aspect most salient for subjects was that their final payment for the entire experiment was "on the line."

8. In addition, our method is superior to the QSR since it is valid beyond the case of risk neutrality.

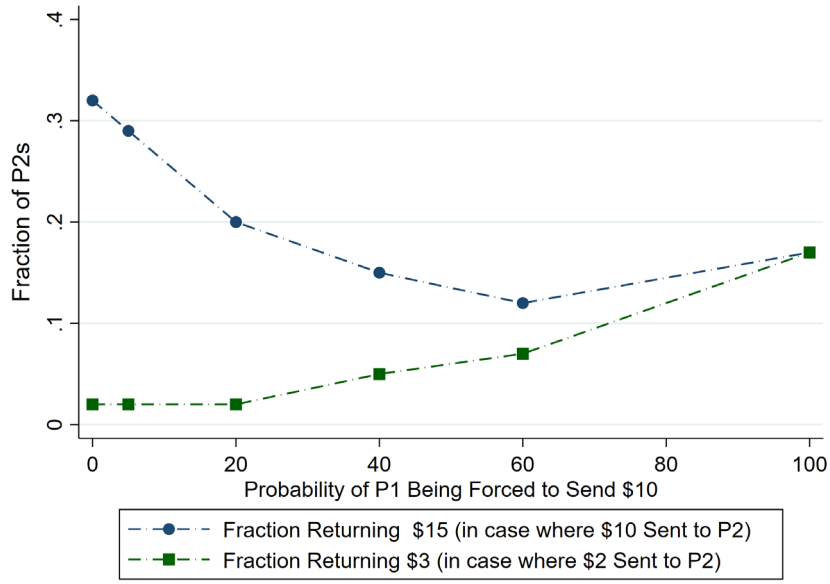
9. See Malmendier, te Velde, and Weber 2014 for a nice summary of internal and external motivations for reciprocity and techniques subjects employ to avoid prosocial behavior.

10. We classify individuals based on their actions taken in the experiment rather than defining types independently. It is plausible that in real-world settings individuals may perform a combination of selfish acts and altruistic acts. It is only when they perform a selfish act that they would need to cover their action. We leave it to future studies to formalize our work.

11. Recall that the payment received under Option 1 is dependent on the action of their opponent. Therefore, the sooner a subject "switches" to Option 1 the higher is their belief about their opponent sending them the high amount.

12. Our definition of "sophisticated" differs from that traditionally used in multiple price list settings. Recall that on the Stage 3 "revealed" belief elicitation form, players are faced with two payment options: Option 1, receiving a payment from the outcome of the game played with their opponent; and Option 2, receiving a payment from an outside gamble. Recall also that for the last decision sheet and corresponding payment option form the chance that the experimenter will force Player 1 to send the whole \$10 to Player 2 is 100%. Therefore, Player 2 will receive \$15 (the highest amount) with 100% probability. Thus, it is in a player's best interest to switch to Option 1 (payment from the outcome of the game) immediately since there is a 100% chance they will receive \$15 from the game, while there is less than a 100% chance they will receive \$15 from the outside gamble.

FIGURE 2
Fraction of Player 2s Returning an Equal Split



necessary to manipulate perceptions of their selfish actions. Examining the behavior of these sophisticated subjects is also necessary to rule out confusion or misunderstanding of the game structure as an explanation of any differences seen between stated and revealed beliefs.

V. MAIN FINDINGS

Eighty-two undergraduate subjects participated in this study, which was conducted at the University of California San Diego, in the Economics Laboratory. Each session lasted about 90 minutes and average earnings were $19 (s.d. $8.16, maximum $32, minimum $10), including a $7 participation fee.

A. *Trust Game Choices*

Result 1a—Selfish Behavior. As we expected the chance that Player 1 is forced by the Experimenter to send $10 increases, the fraction of Player 2s returning $15 (an equal split) declines steadily.

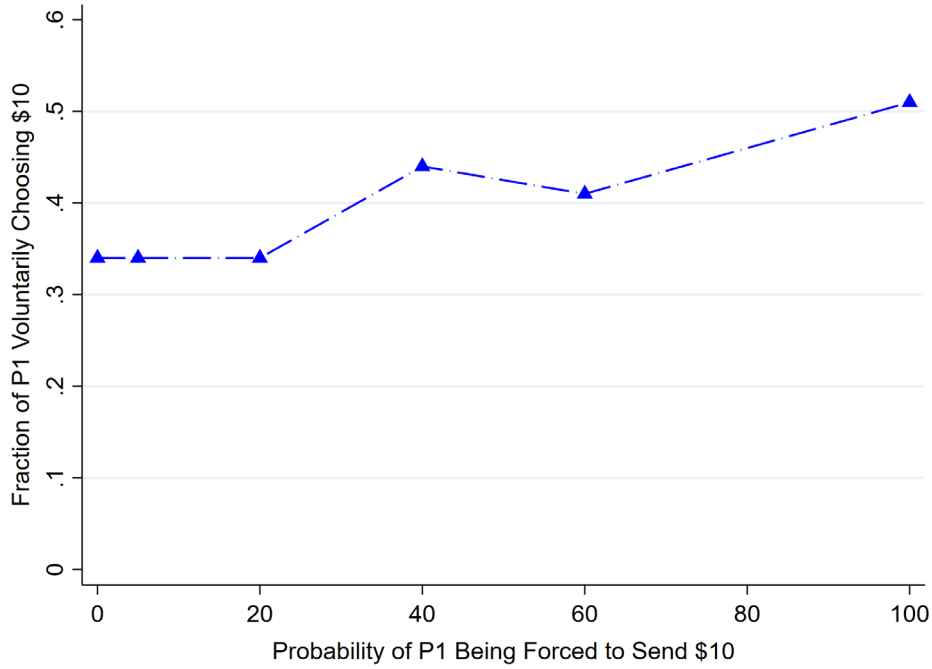
The first column of Table 1 reports the estimates of a random-effects probit model of the probability of Player 2 returning $15 in the case where $10 is sent by Player 1. The second column reports the probability of Player 2 returning $3 in the case where $2 is sent to them. The explanatory variables include indicators for $p \geq .05$, $p \geq .20$, $p \geq .40$, $p \geq .60$, and $p = 1$ (with $p = 0$ omitted). In all cases, we report marginal effects at mean

TABLE 1
Probability of Player 2 Choosing Equal Split, Conditional on Probability of Player 1 Being Forced Random Effects Probit: Marginal Effects

Probability of Player 1 Being Forced to Send $10	(1) If $10 Sent to Player 2: Probability of Player 2 Returning $15	(2) If $2 Sent to Player 2: Probability of Player 2 Returning $3
Constant ($p \geq 0$)	−0.948**	−3.109***
	(0.417)	(0.855)
$p \geq .05$	−0.137	0.000
	(0.393)	(0.751)
$p \geq .20$	−0.717*	−0.210
	(0.422)	(0.817)
$p \geq .40$	−1.105**	0.332
	(0.454)	(0.708)
$p \geq .60$	−1.246***	0.665
	(0.453)	(0.674)
$p = 1$	−0.903**	1.341**
	(0.428)	(0.660)
Observations	246	246

Note: Standard errors in parentheses.
***Significance at $\alpha < .01$; **significance at $\alpha < .05$; *significance at $\alpha < .10$.

FIGURE 3
Fraction of Player 1s Who Sent $10 Voluntarily



values. As we are most interested in Player 2's reaction to knowledge that Player 1 could have been forced to send $10, we focus on the results in the first column. The coefficients in the first column imply that there is a statistically significant decrease in the probability of Player 2 returning $15 when p rises from .05 to .20, from .20 to .40, from .40 to .60, and from $p = .60$ to 1.

We now briefly turn to Player 1 choice behavior. Figure 3 shows the fraction of Player 1s who voluntarily chose to send $10 to Player 2. When the probability of being forced to send the whole $10 to Player 2 is zero, around 30% of Player 1s voluntarily choose to send $10. This fraction gradually increases as the probability that they will be forced to do so increases. Table 2 shows the results from a random effects probit regression. The specification describes the probability of Player 1 voluntarily selecting $10. The explanatory variables include indicators for $p \geq .05$, $p \geq .20$, $p \geq .40$, $p \geq .60$, and $p = 1$ (with $p = 0$ omitted). The coefficients imply that the only statistically significant increase in the probability of voluntarily choosing to send $10 occurs when p rises from .60 to 1 ($\alpha < .10$, one tailed t test).

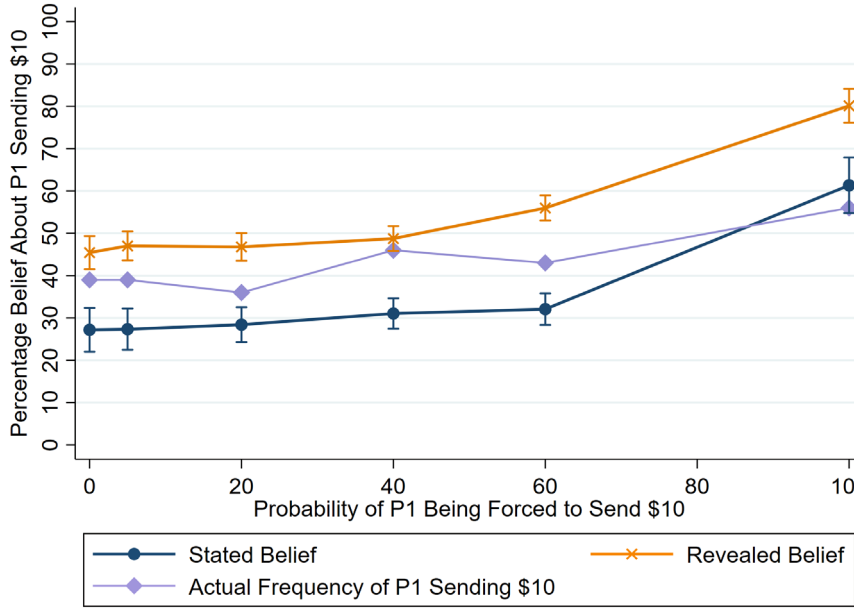
TABLE 2
Probability of Player 1 Voluntarily Sending $10
Random Effects Probit Model

Probability of Player 1 Being Forced to Send $10	Probability of Player 1 Voluntarily Choosing to Send $10
$p \geq 0$	−0.368
	(0.273)
$p \geq .05$	−0.020
	(0.324)
$p \geq .20$	−0.113
	(0.327)
$p \geq .40$	−0.244
	(0.322)
$p \geq .60$	−0.154
	(0.322)
$p = 1$	0.538*
	(0.311)
Mean	0.167
Observations	246

Note: Standard errors in parentheses.
*Significance at $\alpha < .10$.

As can be seen in Figure 2, approximately 30% of Player 2s return $15 to Player 1 when (as was publicly stated) there is zero chance that

FIGURE 4
Player 2 Comparison of Stated vs. Revealed Beliefs



Player 1 was forced to send \$10. In this case ($p = 0$), Player 2s know with certainty that if they receive \$10 that it was Player 1 who decided to send the \$10 and it was of their own volition. Therefore, responsibility for sending the \$10 to Player 2s is unambiguous. However, as the chance that Player 1 will be forced to send \$10 increases, the fraction of Player 2s reciprocating by returning an equal split of \$15 declines. There is a small increase in the number of Player 2s returning \$15 on the last decision sheet, where the probability of Player 1 being forced to send \$10 reaches 100%.

B. Examining Beliefs

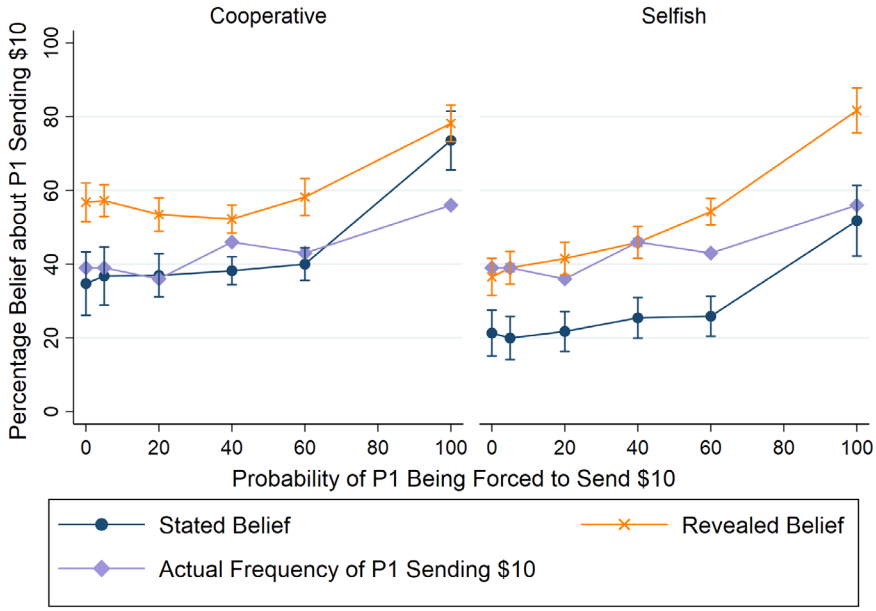
Figure 4 shows Player 2s' stated beliefs, revealed beliefs as well as the actual frequency of Player 1 sending \$10 to Player 2 (recall that this is the probability of Player 1 voluntarily choosing \$10 and not the probability that Player 2 will receive \$10). What is apparent from cursory examination is that there is a constant difference of approximately 20 percentage points between what Player 2s *state* they believe and what Player 2s are *revealed* to believe. This difference is statistically significant for all six values of p.

TABLE 3
Number of Player 2 Types

	Unsophisticated	Sophisticated	Total
Cooperative	10 (24%)	8 (20%)	18 (44%)
Selfish	8 (20%)	15 (36%)	23 (56%)
Total	18 (44%)	23 (56%)	41 (100%)

We begin our exploration of the cause of the difference between stated and revealed beliefs by comparing the stated and revealed beliefs for Selfish and Cooperative Player 2s. We code a Player 2 as being "selfish" if they chose to send \$5 (the lower amount) to Player 1 for every value of p. Otherwise, the subject was coded as "cooperative." Table 3 shows a breakdown of players by type. Looking at Figure 5, it is apparent that both cooperative and selfish Player 2s have a difference between their stated and revealed beliefs. However, the cumulative difference between stated and revealed beliefs is statistically significantly larger for selfish Player 2s (Mann–Whitney $z = 3.089$, $\alpha < .00$). As a further test, we compare Player 2s who have large differences between their revealed and stated beliefs with Player 2s who have little or no difference between what they say they believe

FIGURE 5
Comparison of Player 2 Stated vs. Revealed Beliefs by Player Type



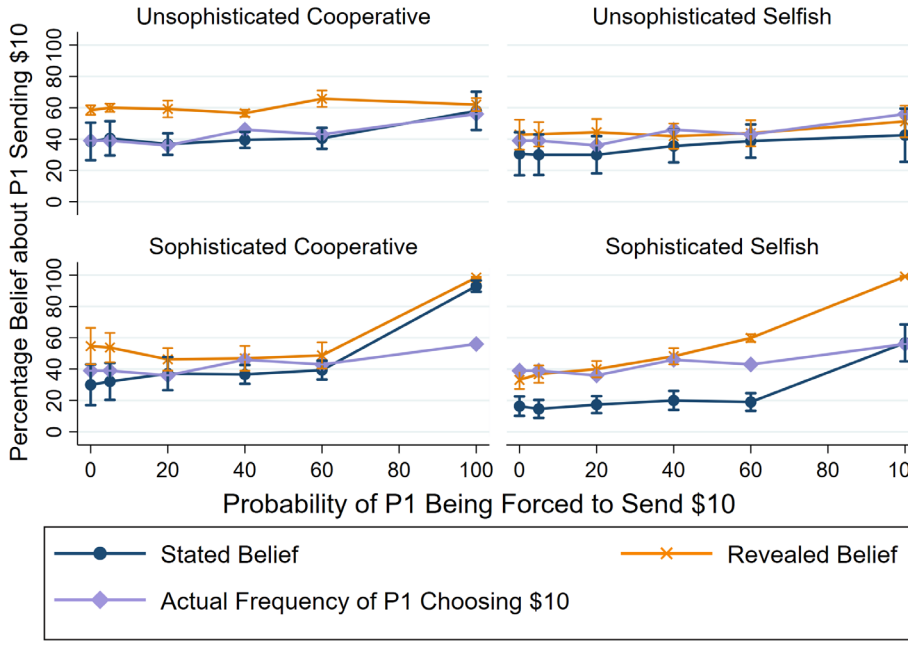
and what they are revealed to believe. Those Player 2s who are "large deviators" are significantly more selfish than those Player 2s with small or no deviations ($t = 4.06$, $\alpha < .00$ two-tailed t test, Mann–Whitney $z = 3.42$, $\alpha < .00$). That is, there appears to be a positive correlation between behaving generously and truthfully stating beliefs. There are three possible explanations for the difference in stated and revealed beliefs. One, that selfish Player 2s were motivated to take the selfish action because they were more pessimistic in their beliefs about Player 1's action than were cooperative Player 2s. Two, the difference is driven by confusion about the game structure. Or three, the difference is driven by subjects who are attempting to excuse selfish behavior by manipulating perceptions about their selfish action. In order to validate our hypothesis, we must be able to distinguish between subjects who are truthful, subjects who are confused, and subjects who are intentionally misstating beliefs in order to manipulate perceptions. To distinguish between these possible explanations, we further explore subject behavior based upon their level of strategic sophistication.

We code a Player 2 as being "Sophisticated" if on the payment option form where $p = 1$, the subject switched from preferring the outside gamble (Option 2) to preferring the outcome of the game (Option 1) in Row 1 or Row 2. Subjects switching in Row 1 or Row 2 of the Payment Option form would have to have reasoned that on the last decision sheet, they were guaranteed to receive $10 as the probability of Player 1 being forced to send $10 was 100% on this sheet. Thus, we assume that subjects who are capable of picking up on this fact understand the game structure better than those who do not.

Result 1b—Perception Manipulation. Sophisticated-selfish Player 2s are *revealed* to believe that there is a much higher chance that Player 1 voluntarily sent $10 than they *state* they believe. Furthermore, the sophisticated-selfish Player 2s are capable of accurately predicting the actual frequency that Player 1 voluntarily chose $10, but when asked, Player 2s state a much lower probability than was true.

Figure 6 shows the stated beliefs, revealed beliefs, and actual frequency of Player 1 voluntarily choosing to send $10 for each type of Player 2 (unsophisticated-cooperative, unsophisticated-selfish, sophisticated-cooperative, and sophisticated-selfish). Comparing the actual frequency line with the revealed belief line, one can see that sophisticated-selfish Player 2s are fully

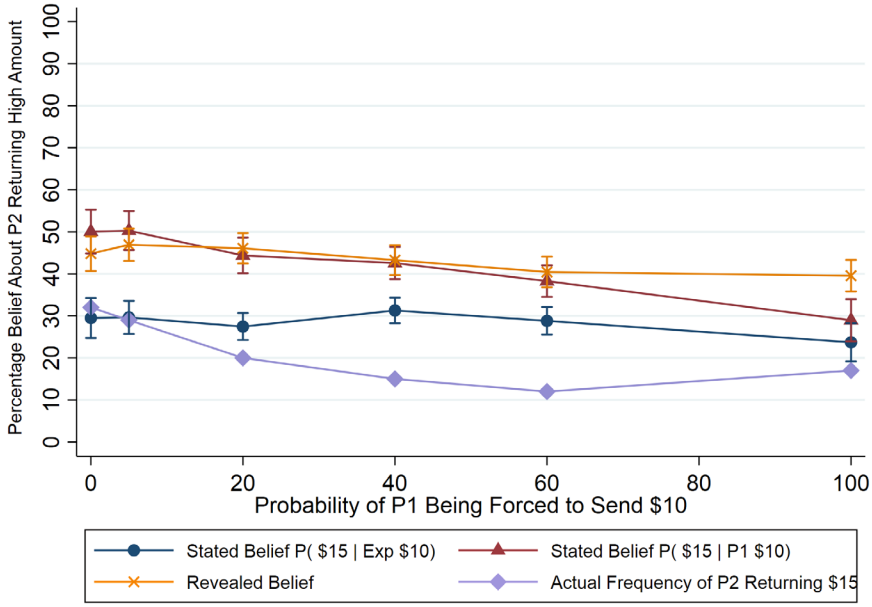
FIGURE 6
Comparison of Player 2 Stated vs. Revealed Beliefs by Player Type



capable of predicting Player 1s' actions. In fact, there is no statistically significant difference between the actual frequency and the revealed belief for $p = 0$, $p = .05$, $p = .20$, and $p = .40$. There is a significant difference for the last two values of p, $p = .60$, and $p = 1$. There is a substantial increase in the revealed beliefs for the last two values of p which causes the difference. The sophisticated-selfish subjects do not best respond to their own stated beliefs, but rather best respond to their revealed beliefs. We can rule out confusion for these subjects as they are strategically sophisticated. We can also rule out pessimism about their opponent, since sophisticated-selfish subjects stake their entire payoff of the game on the chance that their opponent was kind to them—an action they would not take if they truly were pessimistic about the chance their opponent would send them the high amount. This indicates that the sophisticated-selfish subjects intentionally lie about their beliefs when asked, but do not believe their own lies when it comes to the choices they make in the experiment. Unsophisticated-selfish subjects exhibit no statistically significant differences between stated and revealed beliefs, stated beliefs and actual frequency, and occasional significant differences between revealed beliefs and actual frequency. This indicates that unsophisticated-selfish subjects are both honest and realistic in that they truthfully state their beliefs and best respond to these beliefs. While the unsophisticated-selfish have a motive to lie, they seem to either lack the strategic sophistication to carry out the deception or the awareness that such a manipulation is helpful to their social image.

By contrast, sophisticated-cooperative subjects, who have no incentive to hide their cooperative action, exhibit only occasional and small differences between stated and revealed, revealed and actual frequency, and stated and actual frequency. On average, sophisticated-cooperative subjects best respond to their stated beliefs, but sometimes fail to best respond to their stated beliefs in favor of being optimistic. Again, we can rule out confusion for these subjects as they have shown their strategic sophistication. Unsophisticated-cooperative subjects also exhibit a statistically significant difference in stated and revealed beliefs for $p = 0$, $p = .05$, $p = .20$, $p = .40$, $p = .60$, and $p = 1$, which is puzzling since these subjects lack a motive to lie about their beliefs since they were cooperative. However, there is no statistically significant

FIGURE 7
Player 1 Beliefs about Player 2's Actions



difference between unsophisticated-cooperative subjects' stated beliefs and the actual frequency. The unsophisticated-cooperative subjects appear to fail to best respond to their stated beliefs in a way that is overly optimistic. That is, the unsophisticated-cooperative subjects truthfully and accurately state their beliefs when asked, but somehow misunderstood the structure of the revealed belief elicitation. Now to shed further light on whether Player 2s attempt to deceive others, we contrast Player 2 belief data with Player 1 beliefs.

Result 2—Sophisticated Deception. Player 2s are able to accurately anticipate Player 1's expectations of Player 2 behavior which coincide directly with standard social norms of reciprocity. Player 1s both state and reveal that they believe that if Player 2 believes that Player 1 is responsible for voluntarily sending the $10, then Player 2 will reciprocate this kindness by returning $15. Overall, Player 1s expect that if Player 2 believes that Player 1 is undeserving (i.e., sent the low amount) then Player 2 will respond by returning the low amount.

Looking at Figure 7, it can be seen that Player 1s state that they believe that if Player 2 believes Player 1 is responsible for sending the $10 then Player 2 will positively reciprocate.[13]

The pooled group of Player 1s state that they believe that $P(\$15 \mid P1\ \$10) > P(\$15 \mid Exp\ \$10)$ (pooled: $t = 5.79$, $\alpha < .00$, selfish: $t = 4.90$, $\alpha < .00$, cooperative: $t = 3.57$, $\alpha < .00$ two tailed t tests). Now comparing revealed beliefs with stated beliefs, one can see that not only are Player 1s truthful, but are operating on the assumption that Player 2 will positively reciprocate if Player 1 is perceived as responsible for voluntarily sending $10. There is no statistically significant difference between the revealed beliefs and the stated belief of $P(\$15 \mid P1\ \$10)$ for the pooled group of Player 1s. This indicates that Player 1s expect that Player 2 will behave reciprocally, as is the social norm. See Appendixes A and C for additional results.

VI. DISCUSSION OF PERCEPTION MANIPULATION

The discrepancy between the beliefs that selfish-sophisticated players express on their

13. We examine Player 1 behavior in more detail in Appendix C by breaking Pl subjects into two groups, cooperative and selfish Player 1s. Player 1s who chose to send $10 to Player 2 at least three times are coded as "cooperative," otherwise they are coded as "selfish."

stated beliefs and their revealed beliefs raises several questions. First, which belief represents players' "true" beliefs? If selfish-sophisticated players' true beliefs are as they originally stated on their stated beliefs task, then these players are not best responding to these beliefs on their subsequent revealed beliefs task. If they truly believed that chances of Player 1 sending the high amount ($10) were as low as they originally stated on their stated beliefs, then this should have been reflected by their choices on the payment option form in the revealed beliefs task. Instead, their choices on their payment option form indicate that their true underlying belief is that there was in fact a higher chance that Player 1 had sent Player 2 the high amount ($10). Recall that these sophisticated players are the players who were able to correctly calculate and assess that they were better off switching early from Option 2 to Option 1 on Decision Sheet 6, where the chance they would receive the high amount ($10) was 100%. Given that the sophisticated players demonstrated the best strategy in this instance, it is unlikely that they were confused about how to respond on the other parts of the same revealed beliefs task. A further verification of their skill manifests itself in sophisticated-selfish players' ability to accurately predict the actual frequency of their opponents' kindness. By contrast, those players who were selfish, but were unsophisticated, showed more consistency between their stated beliefs and their revealed beliefs. Lack of concern over social image may be one possible explanation for the difference in behavior between the sophisticated and unsophisticated-selfish players. It could also be the case that unsophisticated players do not have enough knowledge to care about how others perceive them, or that they do care about how others perceive them but simply lack the skill to manipulate others' perceptions. The second question that arises is if sophisticated-selfish subjects' revealed beliefs are in fact more representative of their "true" beliefs, then what motivates them to lie on their stated beliefs task? Our explanation for this behavior is that they intentionally misstate their beliefs on the stated beliefs elicitation form in an effort to manipulate how others perceive their selfish actions. Sophisticated-selfish players originally stated that they believed their partner was selfish. To see why this qualifies as perception manipulation let us revisit Player 1 beliefs. Recall that on Figure 6 it was shown that Player 1s stated that they believed that there was a higher chance of positive reciprocity from Player 2 *if Player 2 believed* that Player 1 was personally responsible for sending the high amount. In other words, Player 1 expects that the social norm of reciprocity will be followed by Player 2. Note that the beliefs stated by selfish-sophisticated Player 2s on their stated beliefs form coincide directly with their opponents' (Player 1s') expectations. By stating that they believed they faced a selfish opponent, selfish-sophisticated Player 2s are projecting an image that precisely matches group expectations of their behavior. The fact that selfish-sophisticated Player 2s are so well able to anticipate their opponents' expectations exposes both their keen awareness of and desire to appear to be in compliance with group norms. This statement also provides Player 2 with a ready-made excuse should anyone later inquire about their selfish decision.[14]

A third question that arises is whether stated beliefs differ from revealed beliefs because subjects are forced to "think carefully" about their beliefs in the incentivized revealed belief elicitation. While it is certainly plausible that subjects' beliefs may change over time and with more careful consideration, we only see a large significant difference between stated and revealed beliefs for one group of subjects—the sophisticated-selfish players. If careful consideration were a factor in driving the difference between stated and revealed beliefs, we would expect to see this change across all player types, rather than only in a subgroup of selfish players. Further, if "thinking carefully" could explain the difference we would expect players to both underestimate as well as overestimate their opponents' selfishness. The predominant pattern among the sophisticated selfish is the socially convenient overestimation of their opponents' selfishness.

If one were to look solely at selfish-sophisticated players' actions on their decision sheets or at their stated beliefs alone, it might appear that selfish-sophisticated players had preferences for reciprocity and that it was these preferences that motivated their decision-making process. However, their revealed beliefs paint a different picture entirely. Their revealed beliefs show that these individuals did in truth believe that their partner was kind. In fact, the

14. Concerns for social image maintenance may arise out of a desire to avoid social retaliation or revenge (see Andreoni and Gee 2012 for a review).

sophisticated-selfish players were willing to stake their entire payoff on their belief that their partner was kind. This reveals that individuals are willing to violate the social norms without punishment if they can successfully manage their social image. Thus, behavior previously viewed as supporting hypotheses of reciprocity or guilt-aversion is now shown in a different light. This is not to say that prosocial behavior in the form of pure or impure altruism does not exist. What our results do imply is that if individuals wish to be selfish, attempts to nudge them to cooperate through appeals to reciprocity or guilt may not alter their choice behavior or change their desire to act selfishly. In the face of social pressure to comply with norms, sophisticated selfish individuals may not cooperate but merely take alternative measures to fool others into believing that they are complying with social norms. A fourth and unanswered question is whether the lies of sophisticated-selfish subjects are believed by observers. This remains an open question that deserves more study and is not directly addressed within our paper. Future studies should explore the effects of direct social pressure as well as the effects cognitive ability have on manipulation perception.

VII. CONCLUSION

We explore whether the ability to manipulate perceptions of others' deservingness will eliminate the need to behave altruistically. In order to examine how social pressures affect individuals' decisions to act prosocially, we implement a technique testing whether selfish individuals will lie about the beliefs they held about their opponent when carrying out this selfish action. Specifically, selfish players originally state that they believed their partner would act selfishly toward them. We then administer a second belief elicitation task meant to verify the beliefs stated in the first beliefs task. We find that in the subsequent revealed beliefs elicitation task, selfish players stake their entire payoff on a completely opposite belief than the one they originally stated: that their opponent was in fact cooperative. In order to rule out confusion, we measure the "sophistication" of each subject. Evidence indicates that individuals who are both sophisticated and selfish are the most frequent users of the perception manipulation mechanism. While previous studies of reciprocity concluded that players' beliefs about their opponents' intentions revealed that their subjects had a preference for reciprocity, our results contradict this finding. Our results suggest that individuals have a preference for being *perceived* as being cooperative instead of actually behaving cooperatively. Thus, if one wishes to take a selfish or uncooperative action they can do so without fear of retaliation or punishment so long as they concoct a socially acceptable story justifying their selfish behavior. Given that beliefs seem to be playing a larger role in theory and are being increasingly relied upon as an explanation for behavior, it seems prudent to examine whether expressed beliefs are influenced by social demand to be "socially acceptable."

Our results have important methodological and policy implications. Methodologically, it may be necessary for researchers to exercise caution when designing mechanisms if revealing their true beliefs could potentially interfere with subjects' social image. It is clear from our experiment that even though traditional social pressure measures were not explicitly enacted in the laboratory, subjects bring their habits from their interactions in everyday life into the laboratory.

Previous studies have proposed social pressure or appeals to individuals' emotions as solutions to social dilemmas where tension exists between the group's interests and individual self-interest. However, our results indicate that these aversive incentives may not cause a change underlying selfish desires, but rather instigate a cover-up of the selfish behavior, thereby potentially creating more problems. Social norm compliance may not be enough to encourage prosocial behavior if individuals can appear to comply with the norm without actually behaving in a prosocial manner. We emphasize a need for further study on how to address underlying desires to behave selfishly. Since individuals can easily generate excuses to relieve themselves from social obligations to cooperate, this supports evidence that prosocial behavior can best be motivated by preferences for altruism and/or warm-glow. Further, people who truly want to behave selfishly will do so. Mechanisms designed to apply social pressure or guilt may do nothing to transform selfish behavior to cooperative behavior. Instead, selfish individuals may end up lying in order to maintain the *appearance* that they are cooperating with socially accepted group norms. One might be able to nudge an individual into cooperating *once* in the short term, but as soon as they are able to find a justifying excuse the nudged desired behavior may not be sustained in the long term.

APPENDIX A

TABLE A1
Player 1 Beliefs

	(1) P1s Who Chose $10: Revealed Belief	(2) P1s Who Chose $2: Revealed Belief	(3) All P1s: Revealed Belief	(4) All P1s: Revealed Belief	(5) All P1s: Revealed Belief	(6) All P1s: Revealed Belief
Stated belief about *P*($15/Exp Sent $10)	0.175***	0.141	0.110	0.112		0.138
	(0.064)	(0.137)	(0.117)	(0.114)		(0.090)
Stated belief about *P* ($15 / P1 Sent $10)	0.330***	−0.170	0.125**	0.102*		0.419**
	(0.075)	(0.108)	(0.062)	(0.060)		(0.210)
Stated belief about *P*($3/P1 Sent $2)				−0.143**		−0.695***
				(0.063)		(0.232)
Stated belief about *P*($5/Exp Sent $10)					−0.064	0.043
					(0.112)	(0.066)
Stated belief about *P*($5/P1 Sent $10)					−0.160***	0.302
					(0.059)	(0.209)
Stated belief about *P*($1/P1 Sent $2)					0.074	−0.557***
					(0.072)	(0.207)
Constant	31.173***	38.362***	35.268***	38.227***	51.057***	59.068***
	(5.389)	(2.244)	(3.689)	(3.753)	(11.221)	(6.567)
Observations	97	135	232	232	232	232
Clusters	33	36	39	39	39	19
R-squared	0.278	0.061	0.187	0.060	0.098	0.051

Note: Robust standard errors in parentheses.
***Significance at $\alpha < .001$; **significance at $\alpha < .05$; *significance at $\alpha < .10$.

TABLE A2
Player 2 Revealed vs. Stated Beliefs for Each Value of $P = p_0$ Player 1 Forced to Send $10

***p* = 0**	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	58.5	42.8125	54.688	31.167
Stated belief	38.5	30.625	30	16.333
Difference	20*	12.1875	24.688*	14.834**
***p* = 0.05**	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	60	43.125	53.75	36.833
Stated belief	40.5	30	32.125	14.6
Difference	19.5**	13.125	21.625*	22.233***
***p* = .20**	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	59.25	44.375	46.25	40
Stated belief	36.8	30	37.125	17.333
Difference	22.45***	14.375	9.125	22.667***
***p* = .40**	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	56.5	41.875	46.875	48.214
Stated belief	39.5	35.625	36.625	20
Difference	17***	6.25	10.25	28.214***
***p* = .60**	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	65.75	43.75	48.75	59.833
Stated belief	40.5	38.75	39.375	19
Difference	25.25***	5	9.375	40.833***
***p* = 1**	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	62	51.25	98.438	99.107
Stated belief	58	42.5	93	56.733
Difference	4	8.75	5.438*	42.374***

***Significance at $\alpha < .01$; **significance at $\alpha < .05$; *significance at $\alpha < .10$.

TABLE A3
Player 2 Revealed vs. True Prob Player 1 Voluntarily Chooses \$10 for Each Value of $P = p_0$ Player 1 Forced to Send \$10

$p = 0$	Unsophisticated Nice	Unsophisticated Selfish	Sophisticated Nice	Sophisticated Selfish
Revealed belief	58.5	42.8125	54.688	31.167
True prob Player 1 chooses \$10	39	39	39	39
Difference	19.5***	3.8125	15.688*	−7.833
$p = .05$	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	60	43.125	53.75	36.833
True prob Player 1 chooses \$10	39	39	39	39
Difference	21***	4.125	14.75*	−2.167
$p = .20$	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	59.25	44.375	46.25	40
True prob Player 1 chooses \$10	36	36	36	36
Difference	23.25***	8.375	10.25*	4
$p = .40$	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	56.5	41.875	46.875	48.214
True prob Player 1 chooses \$10	46	46	46	46
Difference	10.5***	−4.125	0.875	2.214
$p = .60$	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	65.75	43.75	48.75	59.833
True prob Player 1 chooses \$10	43	43	43	43
Difference	22.75***	0.75	5.75	16.833***
$p = 1$	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Revealed belief	62	51.25	98.438	99.107
True prob Player 1 chooses \$10	56	56	56	56
Difference	6*	−4.75	42.438***	43.107***

***Significance at $\alpha < .01$; **significance at $\alpha < .05$; *Significance at $\alpha < .10$.

TABLE A4
Player 2 Stated Belief vs. True Probability Player 1 Voluntarily Chooses \$10 for Each Value of $P = p_0$ Player 1 Forced to Send \$10

$p = 0$	Unsophisticated Nice	Unsophisticated Selfish	Sophisticated Nice	Sophisticated Selfish
Stated belief	38.5	30.625	30	16.333
True prob Player 1 chooses \$10	39	39	39	39
Difference	−0.5	−8.375	−9	−22.667***
$p = .05$	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Stated belief	40.5	30	32.125	14.6
True prob Player 1 chooses \$10	39	39	39	39
Difference	1.5	9	−6.875	−24.4***
$p = .20$	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Stated belief	36.8	30	37.125	17.333
True prob Player 1 chooses \$10	39	39	39	39
Difference	−2.2	−9	−1.875	−21.667***
$p = .40$	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Stated belief	39.5	35.625	36.625	20
True prob Player 1 chooses \$10	39	39	39	39
Difference	0.5	−3.375	−2.375	−19***
$p = .60$	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Stated belief	40.5	38.75	39.375	19
True prob Player 1 chooses \$10	39	39	39	39
Difference	1.5	−0.25	0.375*	−20***
$p = .60$	**Unsophisticated Nice**	**Unsophisticated Selfish**	**Sophisticated Nice**	**Sophisticated Selfish**
Stated belief	58	42.5	93	56.733
True prob Player 1 chooses \$10	39	39	39	39
Difference	19	3.5	54***	17.733

***Significance at $\alpha < .01$; **significance at $\alpha < .51$; *significance at $\alpha < .10$.

TABLE A5
Player 1 Stated Belief about $P(\$15 \mid \text{Player 1 } \$10)$ vs. Revealed Belief

	Player 1s Who Chose $2	Player 1s Who Chose $10
$p = 0$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	33.5	65
Revealed	37.28	72.857
Difference	−3.78	−7.857
$p = .05$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	37.8	63.214
Revealed	37.6	72.857
Difference	0.2	−9.643*
$p = .20$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	39.3	58.214
Revealed	36.24	58.929
Difference	3.06	−0.715
$p = .40$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	32.5	55.833
Revealed	33.33	53.333
Difference	−0.83	2.5
$p = .60$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	35.227	47.2
Revealed	35.863	41.412
Difference	−0.636	5.788
$p = 1$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	35.8	42.738
Revealed	25.94	31.429
Difference	9.86	11.309*

***Significance at $\alpha < .01$; **significance at $\alpha < .05$; *significance at $\alpha < .10$.

TABLE A6
Player 1 Stated Belief about $P(\$15 \mid \text{Player 1 } \$10)$ vs. $P(\$15 \mid \text{Exp } \$10)$

	Player 1s Who Chose $2	Player 1s Who Chose $10
$p = 0$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	37.28	72.857
Stated belief $P(\$15 \mid \text{Exp } \$10)$	25.84	36.538
Difference	11.44*	36.319***
$p = .05$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	37.64	72.857
Stated belief $P(\$15 \mid \text{Exp } \$10)$	30.24	28.571
Difference	7.4	44.286***
$p = .20$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	36.24	58.929
Stated belief $P(\$15 \mid \text{Exp } \$10)$	25.24	31.429
Difference	11**	27.5***
$p = .40$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	33.333	53.333
Stated belief $P(\$15 \mid \text{Exp } \$10)$	24.524	39.167
Difference	8.809*	14.166**
$p = .60$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	35.863	41.412
Stated belief $P(\$15 \mid \text{Exp } \$10)$	23.818	35.294
Difference	12.045**	6.118
$p = 1$		
Stated belief $P(\$15 \mid \text{Player 1 } \$10)$	25.941	31.429
Stated belief $P(\$15 \mid \text{Exp } \$10)$	21.389	25.714
Difference	4.552	5.715

***Significance at $\alpha < .01$; **significance at $\alpha < .05$; *significance at $\alpha < .10$.

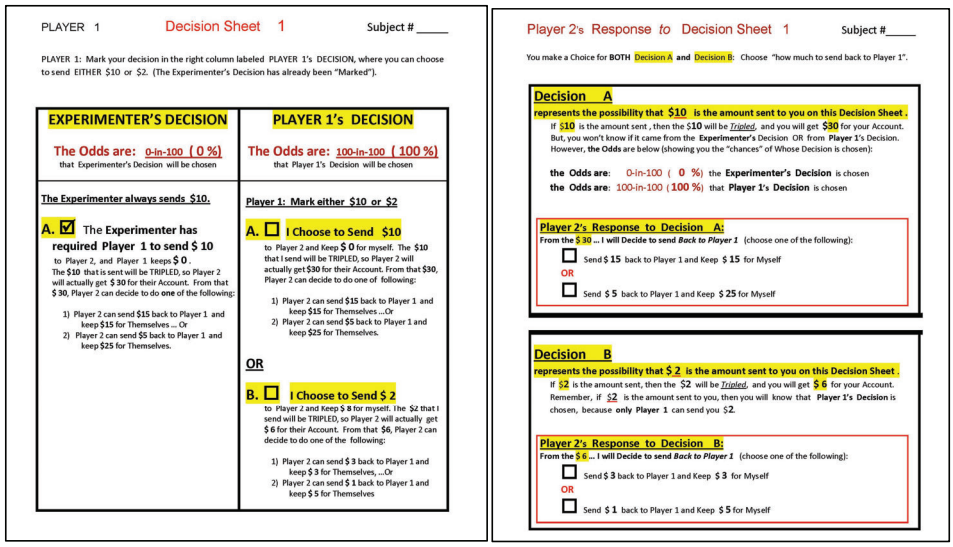
TABLE A7
Player 1 Stated Belief about *P* ($15 | Player 1 $10) vs. Actual Frequency of Player 2 Returning $15

	Player 1s Who Chose $2	Player 1s Who Chose $10
$p = 0$		
Actual frequency of Player 2 returning $15	32	32
Stated belief *P* ($15 | Exp $10)	25.84	36.538
Difference	6.16*	−4.538
$p = .05$		
Actual frequency of Player 2 returning $15	29	29
Stated belief *P* ($15 | Exp $10)	30.24	28.571
Difference	−1.24	0.429
$p = .20$		
Actual frequency of Player 2 returning $15	20	20
Stated belief *P* ($15 | Exp $10)	25.24	31.429
Difference	−5.24*	−11.429**
$p = .40$		
Actual frequency of Player 2 returning $15	15	15
Stated belief *P* ($15 | Exp $10)	24.524	39.167
Difference	−9.524***	−24.167***
$p = .60$		
Actual frequency of Player 2 returning $15	12	12
Stated belief P ($15 | Exp $10)	23.818	35.294
Difference	−11.818***	−23.294***
$p = 1$		
Actual frequency of Player 2 returning $15	17	17
Stated belief *P* ($15 | Exp $10)	21.389	25.714
Difference	−4.389	−8.714*

***Significance at $\alpha < .01$; **significance at $\alpha < .05$; *significance at $\alpha < .10$.

APPENDIX B: SUBJECT FORMS

Stage 1: Decision Sheets



PLAYER 1 Decision Sheet 1 Subject # _____

PLAYER 1: Mark your decision in the right column labeled PLAYER 1's DECISION, where you can choose to send EITHER $10 or $2. (The Experimenter's Decision has already been "Marked").

EXPERIMENTER'S DECISION	PLAYER 1's DECISION
The Odds are: 0-in-100 (0 %) that Experimenter's Decision will be chosen	The Odds are: 100-in-100 (100 %) that Player 1's Decision will be chosen
The Experimenter always sends $10.	Player 1: Mark either $10 or $2
A. ☑ The Experimenter has required Player 1 to send $ 10 to Player 2, and Player 1 keeps $ 0 . The $10 that is sent will be TRIPLED, so Player 2 will actually get $ 30 for their Account. From that $ 30, Player 2 can decide to do one of the following: 1) Player 2 can send $15 back to Player 1 and keep $15 for Themselves ... Or 2) Player 2 can send $5 back to Player 1 and keep $25 for Themselves.	A. ☐ I Choose to Send $10 to Player 2 and Keep $ 0 for myself. The $10 that I send will be TRIPLED, so Player 2 will actually get $30 for their Account. From that $30, Player 2 can decide to do one of following: 1) Player 2 can send $15 back to Player 1 and keep $15 for Themselves ...Or 2) Player 2 can send $5 back to Player 1 and keep $25 for Themselves. OR B. ☐ I Choose to Send $ 2 to Player 2 and Keep $ 8 for myself. The $2 that I send will be TRIPLED, so Player 2 will actually get $ 6 for their Account. From that $6, Player 2 can decide to do one of the following: 1) Player 2 can send $ 3 back to Player 1 and keep $ 3 for Themselves, ...Or 2) Player 2 can send $ 1 back to Player 1 and keep $ 5 for Themselves

Player 2's Response to Decision Sheet 1 Subject #_____

You make a Choice for BOTH Decision A and Decision B: Choose "how much to send back to Player 1".

Decision A
represents the possibility that $10 is the amount sent to you on this Decision Sheet .
If $10 is the amount sent , then the $10 will be *Tripled*, and you will get $30 for your Account.
But, you won't know if it came from the **Experimenter's** Decision OR from **Player 1's** Decision.
However, **the Odds** are below (showing you the "chances" of Whose Decision is chosen):

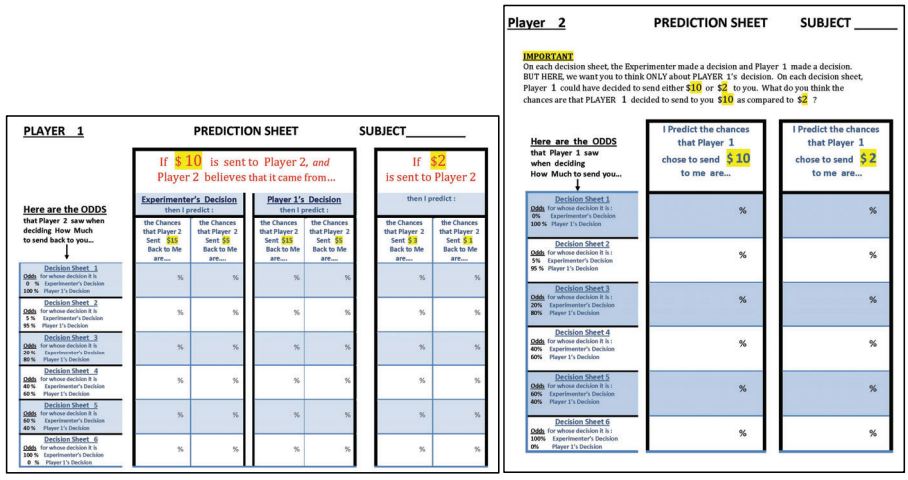
the Odds are: 0-in-100 (0 %) the **Experimenter's Decision** is chosen
the Odds are: 100-in-100 (100 %) that **Player 1's Decision** is chosen

Player 2's Response to Decision A:
From the $ 30 ... I will Decide to send *Back to Player 1* (choose one of the following):
☐ Send $ 15 back to Player 1 and Keep $ 15 for Myself
OR
☐ Send $ 5 back to Player 1 and Keep $ 25 for Myself

Decision B
represents the possibility that $ 2 is the amount sent to you on this Decision Sheet .
If $2 is the amount sent, then the $2 will be *Tripled*, and you will get $ 6 for your Account.
Remember, if $2 is the amount sent to you, then you will know that **Player 1's Decision** is chosen, because **only Player 1** can send you $2.

Player 2's Response to Decision B:
From the $ 6 ... I will Decide to send *Back to Player 1* (choose one of the following):
☐ Send $ 3 back to Player 1 and Keep $ 3 for Myself
OR
☐ Send $ 1 back to Player 1 and Keep $ 5 for Myself

Stage 2: Stated Beliefs ("Prediction Sheet")



PLAYER 1 — **PREDICTION SHEET** — **SUBJECT** _____

Here are the ODDS that Player 2 saw when deciding How Much to send back to you...

	If $10 is sent to Player 2, *and* Player 2 believes that it came from...				If $2 is sent to Player 2	
	Experimenter's Decision (then I predict:)		Player 1's Decision (then I predict:)		(then I predict:)	
	the Chances that Player 2 Sent $10 Back to Me are...	the Chances that Player 2 Sent $5 Back to Me are...	the Chances that Player 2 Sent $10 Back to Me are...	the Chances that Player 2 Sent $5 Back to Me are...	the Chances that Player 2 Sent $2 Back to Me are...	the Chances that Player 2 Sent $1 Back to Me are...
Decision Sheet 1 — Odds for whose decision it is: 0 % Experimenter's Decision, 100 % Player 1's Decision	%	%	%	%	%	%
Decision Sheet 2 — Odds for whose decision it is: 5 % Experimenter's Decision, 95 % Player 1's Decision	%	%	%	%	%	%
Decision Sheet 3 — Odds for whose decision it is: 20 % Experimenter's Decision, 80 % Player 1's Decision	%	%	%	%	%	%
Decision Sheet 4 — Odds for whose decision it is: 40 % Experimenter's Decision, 60 % Player 1's Decision	%	%	%	%	%	%
Decision Sheet 5 — Odds for whose decision it is: 60 % Experimenter's Decision, 40 % Player 1's Decision	%	%	%	%	%	%
Decision Sheet 6 — Odds for whose decision it is: 100 % Experimenter's Decision, 0 % Player 1's Decision	%	%	%	%	%	%

Player 2 — **PREDICTION SHEET** — **SUBJECT** _____

IMPORTANT
On each decision sheet, the Experimenter made a decision and Player 1 made a decision. BUT HERE, we want you to think ONLY about PLAYER 1's decision. On each decision sheet, Player 1 could have decided to send either $10 or $2 to you. What do you think the chances are that PLAYER 1 decided to send to you $10 as compared to $2 ?

Here are the ODDS that Player 1 saw when deciding How Much to send you...	I Predict the chances that Player 1 chose to send $10 to me are...	I Predict the chances that Player 1 chose to send $2 to me are...
Decision Sheet 1 — Odds for whose decision it is: 0% Experimenter's Decision, 100 % Player 1's Decision	%	%
Decision Sheet 2 — Odds for whose decision it is: 5% Experimenter's Decision, 95 % Player 1's Decision	%	%
Decision Sheet 3 — Odds for whose decision it is: 20% Experimenter's Decision, 80% Player 1's Decision	%	%
Decision Sheet 4 — Odds for whose decision it is: 40% Experimenter's Decision, 60% Player 1's Decision	%	%
Decision Sheet 5 — Odds for whose decision it is: 60% Experimenter's Decision, 40% Player 1's Decision	%	%
Decision Sheet 6 — Odds for whose decision it is: 100% Experimenter's Decision, 0% Player 1's Decision	%	%

Stage 3: Revealed Beliefs ("Payment Option Form")

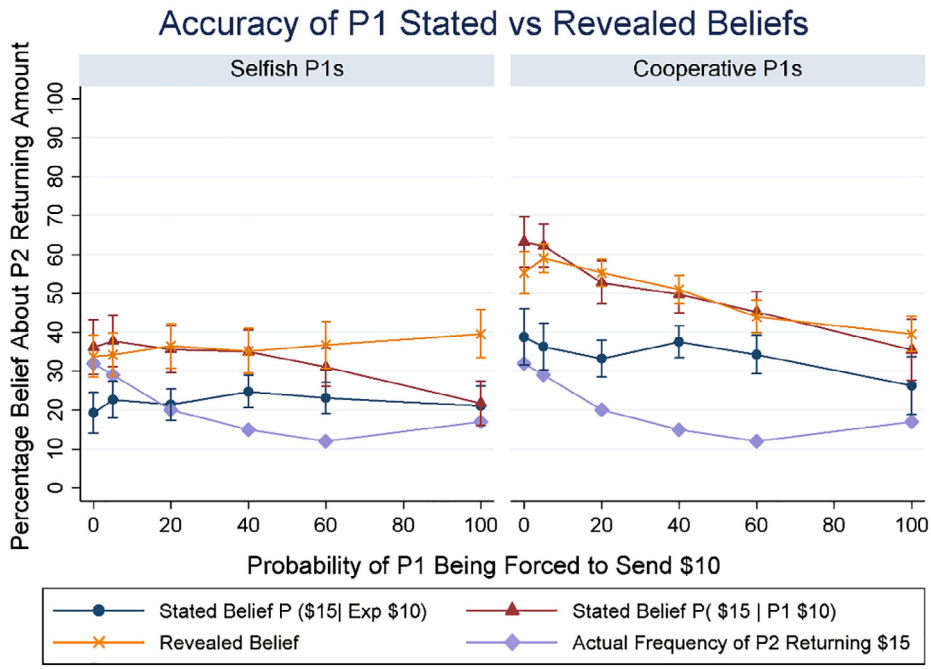
Player 2 — **Payment Option Form** — Subject _____

Option 1		Or	Option 2	
I want to go with the **Outcome of the Game** where I'll get either $5 or $15 depending on what **Player 1** sent me			I would rather go with these Odds here of ...	
			Chance of $5 ↓	Chance of $15 ↓
1)	*Outcome of Game* ☐	*Or*	0 *in* 100	100 *in* 100 ☐
2)	*Outcome of Game* ☐	*Or*	5 *in* 100	95 *in* 100 ☐
3)	*Outcome of Game* ☐	*Or*	10 *in* 100	90 *in* 100 ☐
4)	*Outcome of Game* ☐	*Or*	15 *in* 100	85 *in* 100 ☐
5)	*Outcome of Game* ☐	*Or*	20 *in* 100	80 *in* 100 ☐
6)	*Outcome of Game* ☐	*Or*	25 *in* 100	75 *in* 100 ☐
7)	*Outcome of Game* ☐	*Or*	30 *in* 100	70 *in* 100 ☐
8)	*Outcome of Game* ☐	*Or*	35 *in* 100	65 *in* 100 ☐
9)	*Outcome of Game* ☐	*Or*	40 *in* 100	60 *in* 100 ☐
10)	*Outcome of Game* ☐	*Or*	45 *in* 100	55 *in* 100 ☐
11)	*Outcome of Game* ☐	*Or*	50 *in* 100	50 *in* 100 ☐
12)	*Outcome of Game* ☐	*Or*	55 *in* 100	45 *in* 100 ☐
13)	*Outcome of Game* ☐	*Or*	60 *in* 100	40 *in* 100 ☐
14)	*Outcome of Game* ☐	*Or*	65 *in* 100	35 *in* 100 ☐
15)	*Outcome of Game* ☐	*Or*	70 *in* 100	30 *in* 100 ☐
16)	*Outcome of Game* ☐	*Or*	75 *in* 100	25 *in* 100 ☐
17)	*Outcome of Game* ☐	*Or*	80 *in* 100	20 *in* 100 ☐
18)	*Outcome of Game* ☐	*Or*	85 *in* 100	15 *in* 100 ☐
19)	*Outcome of Game* ☐	*Or*	90 *in* 100	10 *in* 100 ☐
20)	*Outcome of Game* ☐	*Or*	95 *in* 100	5 *in* 100 ☐
21)	*Outcome of Game* ☐	*Or*	100 *in* 100	0 *in* 100 ☐

APPENDIX C: PLAYER 1 BELIEFS

FIGURE C1

Accuracy of Player 1 Stated vs. Revealed Beliefs



Here, we break down Player 1s into selfish and cooperative types based upon their choices in the trust game.

Selfish Player 1s understate their beliefs about P($15 | P1 $10).

The distribution of Player 1s who had the largest difference between their revealed and stated beliefs are significantly more selfish than those Player 1s who had little or no difference between their revealed and stated beliefs ($t = 1.96$, $\alpha < .025$ two tailed t test). Looking at Figure C1, one can see that Player 1s do not exhibit the same degree of deviation from their stated beliefs as exhibited by Player 2s. Also evident from Figure C1 is that "cooperative" player 1s believe there is a higher chance of receiving the high amount back from Player 2 than do the selfish Player 1s. Also, selfish Player 1s were better at predicting the actual frequency that Player 2 would return an equal split than are cooperative Player 1s.

REFERENCES

Akerlof, G. A., and R. E. Kranton. "Economics and Identity." *Quarterly Journal of Economics*, 115(3), 2000, 715–53.

Andreoni, J., and B. D. Bernheim. "Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects." *Econometrica*, 77, 2009, 1607–36.

Andreoni, J., and L. K. Gee. "Gun for Hire: Delegated Enforcement and Peer Punishment in Public Goods Provision." *Journal of Public Economics*, 96(11-12), 2012, 1036–46.

Andreoni, J., and T. Mylovanov. "Diverging Opinions." *American Economic Journal: Microeconomics*, 4(1), 2012, 209–32.

Andreoni, J., and C. Sprenger. "Estimating Time Preferences from Convex Budgets." *American Economic Review*, 102(7), 2012, 3333–56.

Andreoni, J., J. M. Rao, and H. Trachtman. "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving." *Journal of Political Economy*, 125(3), 2017, 625–53.

Ariely, D., A. Bracha, and S. Meier. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review*, 99(1), 2009, 544–45.

Battigalli, P., G. Charness, and M. Dufwenberg. "Deception: The Role of Guilt." *Journal of Economic Behavior & Organization*, 93, 2013, 227–32.

Bénabou, R., and J. Tirole. "Self-Confidence and Personal Motivation." *Quarterly Journal of Economics*, 117(3), 2002, 871–915.

———. "Incentives and Prosocial Behavior." *American Economic Review*, 96(5), 2006, 1652–78.

Berg, J., J. Dickhaut, and K. McCabe. "Trust, Reciprocity, and Social History." *Games and Economic Behavior*, 10(1), 1995, 122–42.

Blount, S. "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes*, 63(2), 1995, 131–44.

Broberg, T., T. Ellingsen, and M. Johannesson. "Is Generosity Involuntary?" *Economics Letters*, 94(1), 2007, 32–7.

Charness, G., and M. Dufwenberg. "Promises and Partnership." *Econometrica*, 74(6), 2006, 1579–601.

Charness, G., and M. Rabin. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*, 117(3), 2002, 817–69.

Dana, J., D. M. Cain, and R. M. Dawes. "What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in Dictator Games." *Organizational Behavior and Human Decision Processes*, 100(2), 2006, 193–201.

Dana, J., R. A. Weber, and J. X. Kuang. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory*, 33(1), 2007, 67–80.

DellaVigna, S., J. A. List, and U. Malmendier. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics*, 127(1), 2012, 1–56.

Di Tella, R., R. Perez-Truglia, A. Babino, and M. Sigman. "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism." *American Economic Review*, 105(11), 2015, 3416–42.

Dufwenberg, M., and U. Gneezy. "Measuring Beliefs in an Experimental Lost Wallet Game." *Games and Economic Behavior*, 30(2), 2000, 163–82.

Falk, A., and U. Fischbacher. "A Theory of Reciprocity." *Games and Economic Behavior*, 54(2), 2006, 293–315.

Fehr, E., and S. Gächter. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14(3), 2000, 159–81.

Harbaugh, W. T. "The Prestige Motive for Making Charitable Transfers." *American Economic Review*, 88(2), 1998, 277–82.

Holt, C. A., and S. Laury, "Risk Aversion and Incentive Effects." *American Economic Review*, 92(5), 2002, 1644–55.

Jacobson, S., and R. Petrie. "Learning from Mistakes: What Do Inconsistent Choices over Risk Tell Us?" *Journal of Risk and Uncertainty*, 38(2), 2009, 143–58.

Lazear, E. P., U. Malmendier, and R. A. Weber. "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied Economics*, 4, 2012, 136–63.

Lisofsky, N., P. Kazzer, H. R. Heekeren, and K. Prehn. "Investigating Socio-Cognitive Processes in Deception: A Quantitative Meta-Analysis of Neuroimaging Studies." *Neuropsychologia*, 61, 2014, 113–22.

Malmendier, U., V. L. te Velde, and R. A. Weber. "Rethinking Reciprocity." *Annual Review of Economics*, 6(1), 2014, 849–74.

Meier, S., and C. Sprenger. "Present-Biased Preferences and Credit Card Borrowing." *American Economic Journal: Applied Economics*, 2, 2010, 193–210.

Rabin, M. "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 1993, 1281–302.

———. "Moral Preferences, Moral Constraints, and Self-Serving Biases." Working Paper No. 95–241, Department of Economics, University of California at Berkeley, 1995.

Schlag, K. H., J. Tremewan, and J. J. Van der Weele. "A Penny for Your Thoughts: A Survey of Methods for Eliciting Beliefs." *Experimental Economics*, 18(3), 2015, 457–90.

Schotter, A., and I. Trevino. "Belief Elicitation in the Laboratory." *Annual Review of Economics*, 6(1), 2014, 103–28.