Self-Signaling Versus Social-Signaling in Giving

Zachary Grossman*

November 4, 2010

Abstract

I investigate the relative importance of social-signaling versus self-signaling in driving giving. I derive specific qualitative predictions about how the response of an image-motivated dictator to a change in the probability that her choice will be implemented depends crucially on the information available to the relevant observer. A probabilistic dictator-game experiment tests the joint, relative, and independent effects of self-signaling and social-signaling. The results provide little evidence of self-signaling, but stronger evidence of social-signaling, particularly in a large subsample that excludes likely 'selfish types'.

Keywords: charitable giving, altruism, dictator game, self-image, self-signaling, signaling, beliefs-dependent preferences

JEL codes: C72,C91,D03,H4

*UC Santa Barbara. Email: grossman@econ.ucsb.edu. I would like to thank Brenda Naputi, Lawrence Sweet and the UC Berkeley Xlab for invaluable support. Matthew Rabin, Botond Kőszegi, Shachar Kariv, Ulrike Malmendier, and Stefano DellaVigna provided helpful comments and criticism. Seminar participants at UC Berkeley, Williams College, American University, UC Santa Barbara, Washington State University, CU Denver, and Claremont-McKenna College provided helpful feedback. This research was supported financially by the Russell Sage Foundation, the UC Berkeley Experimental Social Sciences Laboratory (Xlab), the UC Berkeley Institute for Business and Economic Relations (IBER) and the UC Berkeley Program in Psychology and Economics (PIPE).

1 Introduction

Private giving to individuals and contributions to public goods are major economic activities, yet the motivation behind unselfish behavior it is quite complicated. Giving in experiments is highly sensitive to information about the decision-maker's choice, even in one-shot, anonymous decisions that leave no room for reciprocity.[1] This sensitivity reflects the fact that we care about how our choices affect beliefs, not just outcomes. Social decisions, such as giving, shed important light on socially-valued attributes, such as concern for others or fair-mindedness, that are not directly observable to outsiders and are difficult to introspect. A choice is not only a causal act; it is also an expressive one, sending a signal about the decision-maker's motivation and intentions. As a result, a potential giver may treat the decision as a kind of preference-signaling game, taking into account how her choice will be perceived.

A wealth of psychological evidence points to the fact that people seek to maintain a positive image, both in terms of their self-concept (Bem 1972) and in how they are perceived by others (Goffman 1959). Economists have incorporated concern for image into models of individual decisions by formalizing them as preference-signaling games and by applying the concept of signaling equilibrium to analyze behavior.[2] These models feature a decision-maker with unobservable preferences over outcomes, who also derives value from the endogenously determined beliefs of an observer about those preferences.

Who is the audience for the image-motivated signaler? Whose beliefs matter? When a choice is driven by concern for the beliefs of others, the decision-maker is said to engage in *social-signaling*. Laboratory (Hoffman, McCabe, Shachat, and Smith 1994, Andreoni and Bernheim 2009) and field data (Ariely, Bracha, and Meier 2009, DellaVigna, List, and

[1] For example, while people regularly give in dictator games, many will pay a cost to opt out of the game and prevent the potential recipient from knowing the possibility of the game's existence (Lazear, Malmendier, and Weber 2009, Dana, Cain, and Dawes 2006, Broberg, Ellingsen, and Johannesson 2007). Similarly, many players in the "moral wiggle-room" game of Dana, Weber, and Kuang (2007) choose not to know how their choices affect others, while reverting to selfish behavior.

[2] See, for example, Bernheim (1994), Bodner and Prelec (2003), Benabou and Tirole (2006), and Andreoni and Bernheim (2009).

Malmendier 2009) provide evidence of audience effects and, more specifically, support the hypothesis that giving behavior is consistent with the equilibrium predictions of a Bayesian-rational model of social-image concern.

People also care about their self-image. Bodner and Prelec (2003) argue that because one cannot perfectly introspect the motivation underlying one's own behavior, a person may also distort her behavior in order to manage her impression of herself. Such efforts to maintain positive beliefs about oneself are called *self-signaling*. Benabou and Tirole (2002) justify self-signaling as an attempt to influence the beliefs of a future self who cannot recall the original motivation for the behavior in retrospect.

Directly examining the importance of self-signaling for giving and how it relates to that of social-signaling is necessary for an understanding of unselfish behavior and, as Benabou and Tirole (2006) argue, much like social-signaling, self-signaling is relevant to our understanding how people respond to incentives for prosocial behavior. However, it is difficult to identify self-signaling separately from social-signaling because experimenters cannot manipulate the observability to oneself of one's choice environment and of one's own actions. Thus, previous research has not separately tested for self-signaling in giving.

This paper investigates the relative importance of self-signaling versus social-signaling in driving giving behavior. I also presents a rigorous test of social-signaling and compare the joint, relative, and independent effects of the two kinds of signaling on the frequency of giving, within a unified conceptual framework. An experiment provides the first *direct* test of self-signaling in giving and the first test in any domain of Bayesian-rational self-signaling derived from equilibrium comparative-static predictions.

I write down a simple preference-signaling model of a person who cares about outcomes and the beliefs of an observer. Like that of Benabou and Tirole (2006), the model can be interpreted alternately as one of *social-signaling* or of *self-signaling*, by assuming the observer to be either a separate person or the decision-maker herself, respectively. Applying it to a dictator game in which there is some chance that the dictator's choice will not count, I derive specific predictions about how the qualitative behavioral response to a change in

that probability depends crucially on the information available to the observer.

In an experiment in which participants play such a game, I vary the dictator's choice probability and the information available to the recipient. While the predicted response of a social-signaler to a change in this probability varies with the recipient's information, the dictator always has full information about her choice, so the self-signaling prediction is constant across informational conditions. Holding constant the recipient's information when the choice-probability changes isolates the behavior of a self-signaler.

The results of the self-signaling test provide little support for the hypothesis that self-signaling is a major driver of giving behavior. While support in the overall data is mixed, social-signaling is quite evident in a large subsample, generating swings in giving of over 35 percentage-points. Furthermore, the social-signaling test is quite rigorous, with the data largely conforming to the model's prediction of a particular hierarchy of giving across conditions. Thus, while the influence of the social-image motive on giving behavior is quite clear and largely consistent with rational signaling, the influence of self-image concern on giving, if important, is much more subtle and may involve reasoning and cognitive processes not consistent with a Bayesian signaling model.

Though the motivation is couched in a variety of terms, including 'guilt', 'shame', 'social-image concern', and catch-all terms such as 'audience effects' and 'dislike of not giving', the notion that individuals have preferences over the beliefs of others is well grounded in both theoretically (e.g. Bernheim (1994), Bagwell and Bernheim (1996), Glazer and Konrad (1996)) and experimentally (e.g. Ariely, Bracha, and Meier (2009), Andreoni and Bernheim (2009), Charness and Dufwenberg (2006), Dana, Cain, and Dawes (2006), Tadelis (2008)). Self-image concern, on the other hand, though well established in psychology, has a more recent pedigree in the economics literature.

Psychologists (Bem 1972, Ainslie 1992) have long held that individuals constantly learn and form opinions about themselves and that the internal reward system exerts control over people's behavior is by influencing how they perceive themselves (Baumeister 1998). We cannot always introspect our own preferences and in some contexts we function like an

outside observer, inferring them from our behavior. Kunda (1990) summarizes evidence that people rely on biased cognitive processes to maintain the beliefs that they desire, including positive self-image. However, there is less evidence that individuals distort their actions to modify the flow of information to themselves. Carlsmith and Gross (1969) find that feeling guilty about recent harmful behavior can lead subjects to be more compliant with requests to help an environmental group. Brown and Smart (1991) find that subjects whose self-esteem has been threatened by negative performance feedback on an intellectual task compensate by recruiting positive perceptions of their social qualities, which in turn lead them to behave more prosocially. Thus, while people seek to maintain positive self-image, altruistic acts are the consequence, rather than the instrument. Shaw, Batson, and Todd (1994) find that people try to *avoid* feeling empathy because they understand that that it will lead them to make sacrifices in order to help others.

Bodner and Prelec (2003) argue the relevance of self-signaling for economics and introduce the dual-self modeling approach adopted by others such at Benabou and Tirole (2006) and used herein. However, the only evidence of self-signaling in economically relevant situations is indirect and merely suggestive. Participants in the 'moral wiggle room' game of Dana, Weber, and Kuang (2007) appear to dodge the constraint on selfish behavior imposed by image concerns by avoiding information regarding how their choice affects others. The structure of the decision, as well as the 'plausible deniability' treatment in the same paper and Grossman (2010b) rule out the possibility that the audience is the recipient and suggest that the dictator is motivated to manipulate her own beliefs. However, explaining this behavior with self-signaling requires some degree of self-deception, selective attention, or non-Bayesian thinking, with the dictator ignoring the fact that she chose to avoid the information or delay her choice.[3]

Many studies confound changes to the recipient's information with changes in the dicta-

[3] People desire to perceive themselves as honest even though dishonesty is often profitable. Mazar, Amir, and Ariely (2008) argue that people behave dishonestly enough to profit, but limit their dishonesty so that they may still 'delude themselves of their own integrity" through inattention to moral standards and categorization malleability. While this suggests that individuals distort their behavior so as to maintain their self-concept, it also suggests that the self-image management depends upon self-deception and non-Bayesian rationalizations.

tor's information and choice. Lazear, Malmendier, and Weber (2009) and Broberg, Ellingsen, and Johannesson (2007) find that many people will pay a cost to avoid a chance to share their money if it prevents the potential recipient from learning that the sharing opportunity existed. While avoiding the sharing environment explicitly affects the information of the recipient, it also allows the decision-maker to "avoid" learning how much she would share and is consistent with self-signaling in the manner of Dana, Weber, and Kuang (2007).[4] Testing both signaling phenomena in a unified theoretical and experimental framework allows me to gauge clearly the relative strengths of the effects when they work in opposition. Thus, the experimental results provide insight into the motivation behind behavior in studies in which the two are confounded.

The model, presented in Section 2, follows Bernheim (1994) by taking a game-theoretic approach to an individual decision problem, such as whether or not to give. Like other preference-signaling models, it features a decision-maker with unobservable preferences over outcomes, who also derives value from the endogenously determined beliefs of an observer about those preferences, but it is adapted and simplified for the specific purposes of this study. As in Bodner and Prelec (2003) and Benabou and Tirole (2006), I admit the interpretation of the observer as a dual-self of the self-signaling dictator. Like Tadelis (2008), I consider the effect of the observer's information on the dictator's choice and as in Andreoni and Bernheim (2009) the outcome is probabilistic. Unlike Ellingsen and Johannesson (2008), the utility derived from the observer's esteem is common to all types and does not depend on the audience. Also, in contrast to the continuous models of Benabou and Tirole (2006) and Andreoni and Bernheim (2009), I focus on a binary choice, which simplifies the comparative-static analysis and the experimental design.

The main result characterizes the dictator's response to a drop in the probability that her choice will count for each of three different informational conditions:

- *Observed choice*—the dictator's choice and the probability that it is implemented are

[4] The recipient's beliefs clearly are important. Dana, Cain, and Dawes (2006) show in a similar experiment that costly exit drops significantly when it does not have an explicit affect on the recipient's beliefs.

both observable.

- *Informed*—the observer knows the probability that the dictator's choice is implemented, but does not observe the choice directly, just the outcome fo the dictator game.
- *Not informed*—the observer does not know the probability nor does he observe the dictator's choice. He only observes the outcome.

Across all three conditions, reducing the probability that the dictator's choice will count lowers the expected cost of appearing fair. Under *Observed choice*, this makes the dictator *more* inclined to actually do so. However, when only the outcome is observed, this comes with a commensurate drop in the dictator's power to influence the observer's information, so no net behavioral effect is predicted in the *Not informed* condition. However, when the probability is publicly known, as in the *Informed* condition, a lower probability reduces the impact of a given signal on the observer's beliefs, further reducing the benefit of choosing fairly, thereby causing the dictator to be *less* likely to give.

Furthermore, one can apply the theoretical insights from this paper to reinterpret the results of other studies. For example, Andreoni and Bernheim (2009) vary the dictator's choice probability and focus on how this affects her social-image incentives, while glossing over the fact that this also would affect the behavior of a self-signaller. However, because the dictator's choice probability is public information, but her choice is private, the experimental manipulation parallels that of the *Informed* condition in this paper, in which the self-signaling effect opposes that of social-signaling. Thus, Andreoni and Bernheim's results can be interpreted as all the more solid evidence of social-signaling.

Section 3 presents the experimental design. A dictator chooses between seven units[5] for herself and one for the recipient (7,1), or the more fair allocation of five units apiece (5,5), knowing that with some probability the outcome would be determined by a computer, instead of by her choice. I compare the frequency with which the dictator chooses (5,5) across two probability conditions in which her choice probability is either 1 or 1/3 and in

[5] An experimental currency unit was worth $1.25

three conditions for the recipients information that correspond to those analyzed in the model.

Because a self-signaler always observes her own choice, her giving is predicted to increase as the probability drops, regardless of the recipient's information. In contrast, social-signaling predicts an increase, no change, and a decrease in the frequency of choosing (5,5) in the respective information conditions. Fixing the probability, social-signaling also predicts decreased giving when the perceived signal quality decreases, yielding a complete ranking of giving frequency across all six cells of the 2×3 design. This makes for a strong test of social-signaling, complementing and adding a degree of robustness to previous ones.

The results, presented in Section 4, provide little evidence of self-signaling. When self-signaling is predicted to act in isolation or in conjunction with social-signaling, lowering the probability that the dictator's choice has little impact on the frequency of giving. Furthermore, when the two effects are predicted to act in opposition, the frequency of giving falls significantly, in line with social-signaling and contra self-signaling. When the dictator is certain that her choice will count, improving the recipient's ability to infer her choice from what he observes increases giving from 20% to 35%, evidence of social-signaling.

However, the support for social-signaling in the data is mixed. Despite a large sample, other conditions for which social-signaling predicts effects are not significantly different. Given the strength of previous social-signaling results, this is somewhat surprising, though it might partially be explained by the added rigor of the test.

The average rate of giving across all conditions is quite low and may reflect a subset of participants who are so-called 'selfish-types' or 'money-maximizers' who maximize their own payoff regardless of it's effect on the payoffs or beliefs of others. I use within-subjects data to identify participants likely to fall in this category and find that the evidence of social signaling in a large subsample that excludes these participants is quite compelling. In this group, the ranking of giving frequencies across the six experimental conditions closely matches the hierarchy predicted by the model, ranging from .48 to .11 across the various conditions. Thus, among the large fraction of the population likely to give at all, giving appears to be

largely tied to what it says about the giver—even when the giver is anonymous. Section 5 concludes.

2 A Model of Probabilistic Choice with Beliefs-Concern

A decision-maker (D, female pronouns) plays a binary dictator game. She knows that with some probability the outcome of the game will be determined by chance instead of her choice.[6] She cares directly about the outcome of the game *and* how her choice affects the beliefs of a passive observer (O, male pronouns) about her. I delay discussion of O's identity until the next section, in which I explore the implications of him being the same person as D, as opposed to a distinct person, such as the dictator-game recipient.

Timing First nature draws D's preference type, ρ, from a continuous distribution with full support over the unit interval, and $q > 0$, the probability that D's choice will count. Next, D observes q and O observes $\hat{q}$, a signal of q, described below. Then D and nature simultaneously choose (a_d and a_n, respectively) from the set $\{0, 1\}$, with 1 corresponding to the more fair or generous option in the dictator game. The outcome of the game, a, is then determined. With probability q, the dictator's choice is implemented as the outcome ($a \equiv a_d$) and with probability $1 - q$, nature's choice is implemented ($a \equiv a_n$). Finally, O observes a signal $\hat{a}$, described below, and updates his beliefs about ρ.

Information The distributions from which ρ, q, and a_n are drawn are common knowledge. The three information conditions described in the previous section correspond to the following specifications for the information that O observes:

- *Observed choice*: $\hat{q} = q$ and $\hat{a} = a_d$
- *Informed*: $\hat{q} = q$ and $\hat{a} = a$

[6] The model may be applied to any situation in which a person decides whether or not to commit a costly act while uncertain whether her choice will be implemented.

- *Not informed*: $\hat{q} = \emptyset$ and $\hat{a} = a$

In the *Observed choice* condition, O has full information about D's choice and the conditions under which it is made. In the *Informed* condition, O know the probability that D's choice counts, but only observes the outcome. In the *Not informed* condition, O does not know the realized probability—he only observes the outcome.

Preferences The decision-maker's preferences over the outcome of the dictator game are given by

$$w(a, \rho) = \begin{cases} 0 & \text{if } a = 0 \\ -c(\rho) & \text{if } a = 1 \end{cases},$$

where $c(\rho)$ is the opportunity cost of obtaining $a = 1$. Let $C(\rho) = E[w(a, \rho)|a_d = 0] - E[w(a, \rho|a_d = 1] = qc(\rho)$ denote the opportunity expected cost of *choosing* $a_d = 1$, which decreases (in absolute value) when D's choice is less likely to count. While $c(p)$ may be positive or negative, the crucial sorting assumption of the model is that $c'(\rho) < 0$ and thus, $C'(\rho) < 0$, implying that higher types find it less costly to choose fairly. Thus, ρ captures D's concern for fairness or the well-being of others and w could correspond to any one-dimensional version of standard distributional-preference models such Fehr and Schmidt (1999), Bolton and Ockenfels (2000), or Charness and Rabin (2002).

The decision-maker also cares directly about O's beliefs about her type. Unlike outcome-utility, which is type specific, beliefs-utility is governed by a common valuation function, $v : [0, 1] \to [0, 1]$, a strictly increasing, twice-differentiable function that defines the value of being perceived to be a particular type. Because ρ is private, the decision-maker cares about the expectation of v taken over O's beliefs, which are updated after observing $\hat{a}$ and $\hat{q}$.

An interpretation function, $g : \{0, 1\} \to \Delta([0, 1])$, determines O's posterior beliefs, with $g(\rho, \hat{a})$ denoting the updated probability that D has type ρ conditional on observing $\hat{a}$. This mapping is determined endogenously, but D takes it as given in equilibrium. Thus, D's expected beliefs-utility of a_d, given g, is $V(a_d, g) = E[\int v(r)g(r, \hat{a})\mathrm{d}r|a_d]$, where expectations are taken over the possible realizations of $\hat{a}$ conditional on a_d.

The decision-maker maximizes

$$U(a_d, \rho, g) = E[w(a, \rho)|a_d] + \lambda V(a_d, g),$$

a weighted sum of expected utility derived directly from the outcome (w) and expected utility from beliefs (V). When the weighting parameter, λ, is zero, the model reduces to purely outcome-based social preferences. In this degenerate case, all of the comparative statics derived below for the model and for the experiment predict zero effects. Thus, no purely outcome-based model would generate any of the results detailed below.

Equilibrium Actions and beliefs are determined simultaneously. Equilibrium requires requires all types to maximize utility, taking the observer's interpretation as given, and for the observer's interpretation to be consistent with the action of each type as well as his information about the distributions of a_n and q, and the realized value of q. Formally, an equilibrium consists of an action function, $\sigma : [0, 1] \to \{0, 1\}$, and an interpretation function, g, for which

- for each type ρ, $U(\sigma(\rho), \rho, g) \geq U(\sigma'(\rho), \rho, g)$ for any $\sigma' : [0, 1] \to \{0, 1\}$ and
- g follows Bayes rule, when appropriate.

Applying the D1 refinement (Cho and Kreps 1987, Banks and Sobel 1987) eliminates pooling equilibria based upon unreasonable beliefs, and I further refine equilibrium by eliminating those that are unstable in the sense that small deviations from equilibrium induce disequilibrium incentives that drive the system further from it.

I restrict attention to pure strategies because the sorting assumption requires that at most one type is indifferent in equilibrium. This guarantees that, in equilibrium, the fair allocation is chosen by higher types and therefore confers more favorable beliefs, which is stated formally as Fact 1.

Fact 1 (Monotonicity)**.** *In equilibrium, if* $\sigma(\rho_1) = 0$ *and* $\sigma(\rho_2) = 1$*, then* $\rho_1 < \rho_2$ *and* $V(0, g) < V(1, g)$.[7]

[7] All proofs are presented in Appendix A.

For non-pooling equilibria, monotonicity allows one to restrict attention to strategies and beliefs that can be characterized by a cutoff ρ^*, above which all types choose $a_d = 1$ and below which all types choose $a_d = 0$. Let $V(a_d, \rho)$ denote the expected beliefs utility from choosing a_D when the observer's beliefs are characterized by $\rho^* = \rho$. The cutoff type's indifference condition can be written as

$$C(\rho^*) = \lambda B(\rho^*),$$

where $B(\rho) = V(1, \rho) - V(0, \rho)$ is the beliefs-utility benefit of choosing $a_d = 1$ *when the cutoff is* ρ. This benefit is the same for all types and is a function of the cutoff. Monotonicity guarantees that $B(\rho)$ is strictly positive, which implies that the cutoff type strictly prefers (in terms of outcome-utility) the less-fair outcome.

Comparative Statics The propositions below form the basis of the experimental tests described in Section 3. To facilitate stating them, I introduce k, which denotes O's belief—prior to observing $\hat{a}$—about the probability that $\hat{a}$ will equal a_d. Using the available information, O derives the value of k according to Bayes rule:

- $k = 1$ in the *Observed choice* condition, because $\hat{a} = a_d$
- $k = \hat{q} = q$ in the *Informed* condition, because $\hat{q} = q$, but $\hat{a} = a$
- $k = E[q]$ in the *Not informed* condition, because $\hat{q} = \emptyset$ and $\hat{a} = a$

The effect on k of a change in q depends on the information condition. When O is *Informed*, k changes in lockstep with q. However, when O is *Not informed*, he does not know the realization of q, so k is independent of q. Similarly, in the *Observed choice* condition, O knows a_d regardless of q's value, so k is fixed at 1.

Let $B^1(\rho)$ denote the specific form taken by $B(\rho)$ when D's choice is observable (so $k = 1$). While this baseline benefit function depends only on the equilibrium cutoff, Lemma 1 states that when a_d is not directly observable, $B(\rho)$ also depends on q and k. Specifically, the beliefs-utility benefit diminishes when D's choice is less likely to actually influence O's observation or when O's signal has lower quality.

Lemma 1. *When a_d is not observable, $B(\rho) = qkB^1(\rho)$.*

Proposition 1 characterizes the behavioral effect of changing the realized probability that D's choice counts.

Proposition 1. *1.* $\frac{\partial \rho^*}{\partial q} > 0$ *in the* Observed choice *condition;*

2. $\frac{\partial \rho^*}{\partial q} = 0$ *in the* Not informed *condition;*

3. $\frac{\partial \rho^*}{\partial q} \leq 0$ *in the* Informed *condition, with equality if and only if* $q = 0$.

The effect of changing q varies sharply across the three information conditions. In all three conditions, lowering the probability that D can influence the outcome has the same effect on the left-hand side of the indifference equation. It cheapens the expected outcome-utility cost of choosing fairly (which is positive near the cutoff), which in isolation would induce more types to do so. In the *Observed choice* condition, this accounts for the entire net effect because the beliefs-utility benefit *function* depends only on ρ^*.

However, in the *Not informed* condition, q reflects both D's ability to influence the outcome and her ability to influence O's observation. Thus, any change in q affects both sides of the indifference equation in the same manner. A drop in the outcome-utility cost of choosing fairly is exactly matched by a drop in the beliefs-utility benefit, leading to zero net effect. In the *Informed* condition, lowering q also lowers the quality of $\hat{a}$ as a signal for a_d, leading O to further discount his observation. This *further* lowers the beliefs-utility benefit, leading to *less* than the original amount of giving.[8]

[8] Shocks to C such as rewards or incentives, as well as the changes in q considered here *can* undermine the equilibrium beliefs-utility benefit of choosing fairly, an example of the overjustification effect (Lepper, Greene, and Nisbett 1973). However, unlike the model of Benabou and Tirole (2006), motivational crowding out can not occur in *net*. Benabou and Tirole (2006)'s backfiring-incentives result hinges upon two dimensions of uncertainty, and in this one-dimensional model the conditions under which net crowding out can occur are precisely those that render equilibrium unstable. See Appendix B for a discussion of stability and a characterization of necessary and sufficient conditions for stability.

Assuming unidimensional preferences might not be reasonable for situations in which prosocial behavior requires time or effort (such as giving blood) and may also be accompanied by a monetary reward. However, in typical experimental dictator games fairness and monetary preferences can more reasonably be captured in a single parameter. Setting aside the direction of the effect, the mere fact that behavior is sensitive to q cannot come out of a distributions-based model.

While k cannot change independently of q in a given information condition, fixing q, the value of k by definition depends on O's information. Proposition 2 characterizes the qualitative effect of a change in k independent of q, namely that increasing k will lead more types to choose fairly.

Proposition 2. $\frac{\partial \rho^*}{\partial k} < 0$

A higher k means that that O's observation has a greater impact on his beliefs, increasing the beliefs-utility benefit of choosing fairly without affecting the outcome-utility cost. This formalizes the effect observed by Dana, Weber, and Kuang (2007), whereby obscuring the link between actions and their consequences (in terms of beliefs) diminishes fair behavior.

3 Experimental Design

Participants played a probabilistic dictator game during one of twenty-four experimental sessions, each with 8-24 participants, conducted at the Experimental Social Science Laboratory (XLab) at the University of California-Berkeley in July through October 2007. The 379 participants were drawn from a pool of university students and staff and the sessions lasted approximately one hour. Payoffs were stated in terms of experimental currency units worth \$1.25 each and the average earnings were around \$19.

The participants sat at desks separated by privacy dividers and arranged in four parallel rows. They read instructions and communicated decisions by computer using Z-Tree (Fischbacher 2007). So as to limit the role of the observer to either the recipient or the dictator herself, as opposed to the experimenter, participants' identities and choices were unknown to the experimenter. This anonymity was emphasized in the instructions, which are reproduced in Appendix D. Participants faced a total of five decisions, however this paper focuses on and reports the result of only one, with a second reported in Appendix C. The excluded decisions are not relevant to the current research question and their results do not directly contradict any of the results or conclusions of this paper.[9]

[9] In addition to the probabilistic dictator game reported, participants also played one in which the payoffs were

The structure and timing of the game was as described in Section 2, with the following additional details. Each participant played the role of dictator and served as recipient for someone else, with the total payoff for the decision being the sum of the payoff for each role. The two allocations were (7,1) and (5,5), where the first number indicates the dictator's payoff (in experimental currency units) and the second indicates the recipient's payoff. 'Nature's' choice was determined by computer. For half of the dictator-recipient pairs it was (7,1) with probability 1 and for the other half it was (5,5) with probability 1.[10] The uncertainty as to whether the dictator's choice would be implemented was also realized by the computer.

The treatments followed a 2×3 design. With equal probability, dictators were assigned to either the *High* (H) probability condition, in which $q = 1$, or the *Low* (L) probability condition, in which $q = 1/3$. Each recipient was assigned to one of the informational conditions described in Section 2: *Observed choice* (O), *Not informed* (N), or *Informed* (I) and the matching dictator was informed of this assignment.[11] Matchings were made within each informational condition so that subjects would not be aware of the informational manipulation and would only have to learn the instructions for one condition.

Applying the model to the experiment, I make two crucial assumptions. First, because the (5,5) outcome is more equitable, features a higher minimum payoff, and greater total payoff, I assume it to be the fair outcome.[12] Second, the dictator can always observe her own choice. Thus, while the recipient's information varies across conditions, the dictator always has the same information.

The model from Section 2 may be interpreted as featuring either social-signaling or self-

(7.50,3.75) and (4,4). These decisions were presented simultaneously and the order in which the decisions and payoffs were displayed was randomized. The second set of payoffs was chosen specifically because different fairness criteria (equity, maximin, efficiency) disagree about which outcome is more fair and is not suitable for addressing the current research question. One of these two decisions was selected randomly for payment. The other three decisions were presented sequentially and one of these was randomly selected for payment. All matchings were random and anonymous, with separate matchings for each decision.

[10] The computer chose each outcome with equal probability and both players were told the realization of this choice in the instructions.

[11] In earlier sessions subjects were assigned to the I or N conditions with equal probability, while in later sessions subjects were twice as likely to be assigned to the N condition. The O condition was run in the final four sessions.

[12] Any standard distributions-based model of social preferences would agree with this assessment and would therefore lead to the same comparative static predictions were it to be used as w.

signaling, depending upon the identity of the Observer. The social-signaling interpretation of the model puts the recipient in the role of the Observer, so the experimental information condition corresponds to that of the Observer. In the self-signaling interpretation, however, the Observer is the dictator herself, so the Observer is always in the *Observed choice* condition regardless of the *experimental* information condition.

The first two predictions spell out how the practical implications of Proposition 1 depend upon whether the model is viewed through the lens of self-signaling or social-signaling. Prediction 1 states that that in *any* given information condition (in the experiment) a self-signaling dictator whose choice counts with *Low* probability is more likely to choose fairly than one facing *High* probability.

Prediction 1 (Self-signaling)**.** *The proportion of subjects choosing (5,5) will have the following rankings across probability conditions:*

$$OH < OL, \qquad NH < NL, \text{ and} \qquad IH < IL,$$

where the first letter in each pair indicates the recipient's informational condition and the second letter indicates the dictator's probability condition.

In contrast, the response of social-signaling dictators varies with the information condition, in accordance with the three parts of Proposition 1.

Prediction 2 (Social-signaling)**.** *The proportion of subjects choosing (5,5) will have the following ranking across probability conditions:*

$$OH < OL, \qquad NH = NL, \text{ and} \qquad IH > IL.$$

Thus, the three respective information conditions offer insight into the joint, relative, and independent effects due to self- and social-signaling. In *Observed choice* the effects work in tandem, in *Informed* they are opposed, and in *Not informed* only self-signaling predicts an effect, thereby identifying the effect of self-signaling independent of social-signaling.

The next two predictions rely on Proposition 2, which applies to comparisons across informational conditions while holding constant the probability. Prediction 3 states that for a given probability condition, the behavior of a (purely) self-signaling dictator will not vary across (experimental) information conditions.

Prediction 3 (Self-signaling)**.** *In a given probability condition, the proportion of subjects choosing (5,5) will be equal across information conditions.*

In contrast, according to Prediction 4 the social-signaler's behavior varies. In the IH condition, $k = q = 1$ and thus it is structurally equivalent to OH, with behavior predicted to be the same. The recipient's signal is more noisy in the NH condition, however, so less giving is expected. Similarly, for *Low* probability dictators, giving is predicted to fall as k decreases from the OL to the NL to the IL condition.

Prediction 4 (Social-signaling)**.** *The proportion of subjects choosing (5,5) will have the following rankings across information conditions:*

$$NH < IH = OH \text{ and } IL < NL < OL.$$

Together, these two predictions present constant-probability comparisons across information conditions as a way to identify the effect of social-signaling independent of self-signaling. Finally, combining Predictions 2 and 4 yields the complete hierarchy of giving across conditions predicted by social-signaling (in isolation).

Prediction 5 (Social-signaling)**.** *The proportion of subjects choosing (5,5) will have the following ranking across all conditions:*

$$IL < NL = NH < IH = OH < OL.$$

The vast literature on dictator games shows that, for many individuals, maximizing the one's monetary payoff is not the exclusive objective. However, across hundreds of variations of dictator games, manipulating a diversity of factors (e.g framing, social-distance, price,

stakes, and blindness), a small, yet non-trivial share of participants (typically 15 to 35 percent) does choose the outcome that maximizes the dictator's monetary payoff. For example, in the 31 experimental conditions included in Camerer (2003)'s summary of allocations in dictator games (Table 2.4), 33% of the 1042 participants chose the money-maximizing allocation. Experimental studies of player heterogeneity typically find one-fifth to one-third of the population to be "selfish types" (e.g. Fischbacher, Gächter, and Fehr (2001), Kurzban and Hauser (2005), Burlando and Guala (2004)).

Given the presence of money-maximizers, the model will not accurately predict the behavior of all participants. To investigate the extent to which signaling effects exists within a sizeable subset of the population, I repeat the data analysis on two subsamples that exclude participants categorized as likely to be money-maximizers using data from a separate dictator game played in the same session.

In this second dictator game, the degree of intra-subject anonymity varied across two conditions. Varying anonymity, and thus the recipient's ability to identify the dictator, while holding constant the recipient's information about the dictator's choice should not affect the behavior of a self-signaler. Moreover, while it may affect the behavior of a dictator who cares about her social-image, the model does not directly address uncertainty as to the dictator's identity (though it could easily be adapted to do so) and relaxing anonymity introduces the specter of post-experiment retaliation. Thus, this decision is included only to identify potential money-maximizers, not because it directly sheds light on the predictions of the model.

Briefly, in this game dictators chose an amount, $t \geq 0$, to transfer from their endowment to an anonymous recipient. In the *High Anonymity* condition, recipients have no information about the identity of the recipient, while in the *Low Anonymity* condition the recipient is told the row in which the dictator is seated, thereby lowering the degree of anonymity by a factor of four. A more detailed description can be found in Appendix C.

4 Results

4.1 Analysis of Overall Data

Of the 379 participants in the probabilistic dictator game, 97 (26%) chose (5,5). Table 1 summarizes the frequency of choosing (5,5) by condition, with the first two columns presenting the overall data. There is little variation across conditions in the frequency of giving. A chi-square test ($\chi^2(5) = 7.82$, $p < .17$) cannot reject the hypothesis that the probability of giving is the same in all conditions, which is consistent with the hypothesis that behavior is primarily driven by distributional concerns. In light of previous evidence contradicting this hypothesis and directly supporting social-signaling, this is rather surprising and might be attributed to noise introduced by some of the features of the experiment, such as multiple decisions, doubling up roles, and complicated instructions delivered exclusively by computer.

Table 1: Frequency of fair choice (5,5) in probabilistic dictator game (sample size)

	Overall data		Small Subsample		Large Subsample	
Info.\Prob.	*High*	*Low*	*High*	*Low*	*High*	*Low*
Observed choice	.35 (40)	.25 (40)	.33 (21)	.47 (15)	.33 (21)	.47 (15)
Informed	.36 (58)	.20 (56)	.78 (9)	.08 (12)	.48 (23)	.11 (27)
Not informed	.20 (89)	.24 (96)	.15 (13)	.29 (17)	.18 (34)	.28 (36)

Consider the specific predictions based on Proposition 1, beginning with the self-signaling test. Among dictators with *Not informed* recipients, 20% chose (5,5) in the *High* condition, compared to 24% in the *Low* condition. Though the direction of change is as predicted, the difference in proportions is not statistically significant ($z = .61$, $p < .27$)[13] so the direct test offers little support for self-signaling. Moreover, with a sample of almost 200 participants, the test is quite powerful. Any proportion over .31 in the *Low* condition would have been sufficient to reject at the 5% level the null hypothesis of no self-signaling.

The remaining two conditions offer no support for self-signaling and mixed support for

[13] Hypothesis tests are one-tailed z-tests of proportions with pooled samples, unless otherwise noted.

social-signaling. Giving drops from the *High* to *Low* conditions, both in *Observed choice* (35% to 25%) and *Informed* (36% to 20%), which in both cases is inconsistent with self-signaling. While the drop in *Observed choice* is not consistent with social-signaling, the difference is not statistically significant ($z = 0.98$, $p < .17$). On the other hand, in the *Informed* condition the drop *is* consistent with social-signaling, and the difference is significant ($z = 1.97$, $p < .03$).

The comparisons based on Proposition 2 offer further mixed support for social-signaling. As predicted, when the choice probability is *High*, the rate of fair behavior is virtually identical in the *Observed choice* (35%) and *Informed* (36%) conditions, and exceeds that in the *Not informed* condition by a significant margin, whether considered separately ($z = 1.80$, $p < .04$ for OH vs. NH; $z = 2.14$, $p < .02$ for IH vs. NH) or pooled ($z = 2.34$, $p < .01$). However, among the dictators in *Low*, the proportions giving in the *Informed*, *Not informed*, and *Observed choice* conditions are .2, .24, and .25 respectively. While this conforms to the prediction that giving will increase monotonically across these conditions, the proportions are not significantly different from each other and in fact are strikingly similar.

4.2 Analysis of Subsamples that Exclude Money-Maximizers

Two hundred and one of the 379 participants played the second dictator game, with the complete results presented in Appendix C. While 36 out of 69 (52%) dictators in the *High Anonymity* condition chose to transfer $t = 0$, in the *Low Anonymity* condition only 45 out of 132 (34%) did so. This group of 45 participants, unswayed to give by concern for the payoffs of others, self-image, and—with limited anonymity—the beliefs and potential responses of others, is more likely to have a high concentration of money-maximizers than the group of non-givers from the *High Anonymity* condition. To define the first subsample, I exclude these 45 participants and retain the 87 dictators in the *Low Anonymity* condition that gave a positive amount to the recipient.

The third and fourth columns of Table 1 show the probabilistic dictator game results for

this subsample. The most striking feature of the results is the increased variation in giving rates across conditions, relative to the overall data, with the values ranging from 77.8% in the *IH* condition to 8.3% in the *IL* condition. While a smaller sample can be expected to be noisier, a chi-square test rejects the hypothesis that the giving rate is the same across all conditions ($\chi^2(5) = 14.58$, $p < 0.012$), though only at the 7% level with Yates' correction ($\chi^2(5) = 10.37$, $p < 0.065$). Furthermore, much of the variation provides strong support for social-signaling, though none of the differences predicted by self-signaling are significant.

As in the overall sample, the difference in proportions giving across probability conditions is significant in the *Informed* condition ($z = 3.24$, $p < .001$), but not in the *Not informed* ($z = .55$, $p < .29$) or *Observed choice* condition ($z = .81$, $p < .21$). At almost 70 percentage points, the difference in the *Informed* condition is quite striking. The evidence for self-signaling is at best weak, though it is stronger than in the overall sample. Specifically, the *Observed choice* and *Not informed* differences are greater than in the overall sample (and in the appropriate direction).

Comparing *across* information conditions, the social-signaling predictions of Proposition 2 are largely confirmed. Among *Low* dictators, giving increased monotonically from the *Informed* (8%) to the *Not informed* (29%) to the *Observed choice* (47%) condition, as predicted. Unlike in the overall sample, the difference between the two extreme conditions is statistically significant ($z = 2.17$, $p < .02$).

Among *High* dictators, 15% chose fairly in the *Not informed* condition, while 78% and 33% did so in the *Informed* and *Observed choice* conditions, respectively. The differences in proportions are statistically significant ($z = 2.93$, $p < .002$ and $z = 1.95$, $p < .03$, respectively) when *Not informed* is compared to either *Informed* or *Informed* and *Observed choice* pooled together, though not compared to *Observed choice* alone ($z = 1.15$, $p < .13$). Although theory predicts the same level of giving in the *OH* and *IH* conditions, the difference of 45% is statistically significant ($z = 2.24$, $p < .03$).

Restricting attention to subjects choosing $t > 0$ in the *Low Anonymity* condition severely reduces the size of the sample. To assuage doubts about the validity of the hypothesis test

on such a small sample, I examine a second subsample that enlarges the first by including *all* 69 subjects in the *High Anonymity* condition. Of the 201 dictators in the second game, this larger subsample excludes only the 45 (22.3%) who gave nothing in the *Low anonymity* condition.

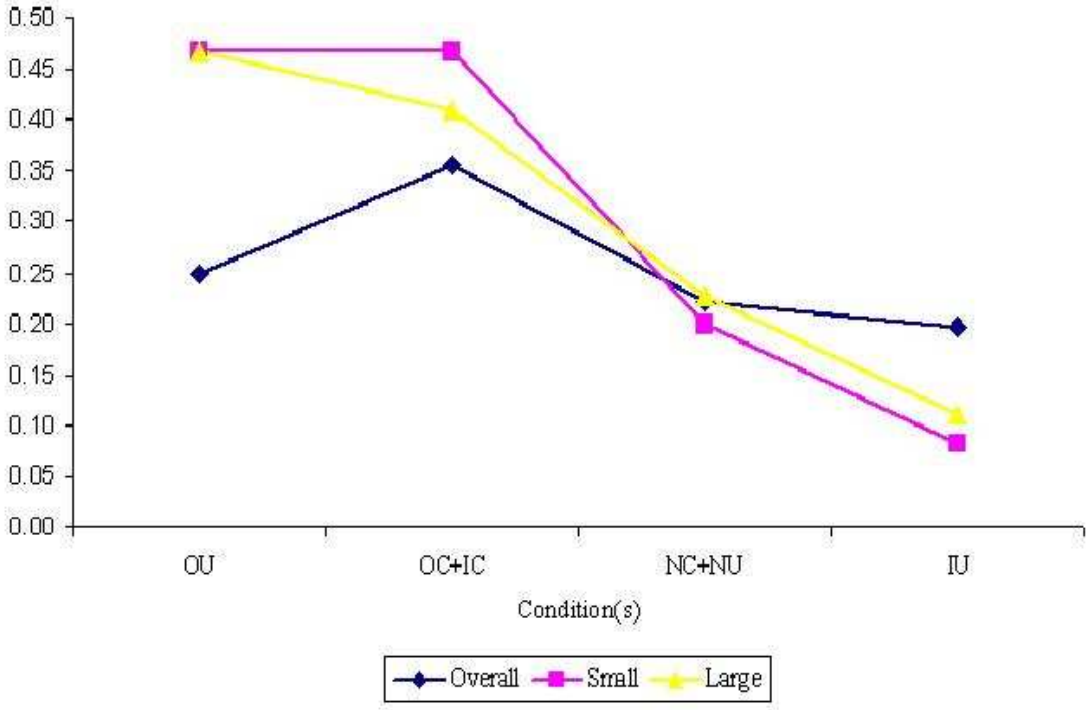
The probabilistic dictator game results for the larger subsample, shown in the last two columns of Table 1, show that the strong social-signaling effects found in the smaller subsample are robust, and persist in a large fraction of the population. Again there is little evidence of self-signaling: lowering the dictator's choice probability in the *Not informed* condition increases the frequency of choosing (5,5) from .17 to .28 ($z = 1.01$, $p < .16$), but the difference is not statistically significant. However, this is consistent with Prediction 2, as is the drop in giving from 48% to 11% ($z = 2.88$, $p < .002$) in the *Informed* condition.[14]

In line with Prediction 4, when the dictator is *High*, the giving frequencies in the *Observed choice* (.33) and *Informed* (.48) conditions are not significantly different ($z = 0.98$, $p < 0.33$), and the .18 in the *Not informed* condition is significantly different from *Informed* ($z = 2.44$, $p < .01$), *Observed choice* ($z = 1.33$, $p < .10$), and the pooling of those two samples ($z = 2.21$, $p < .02$). For *Low* dictators, the giving increases from 11% in the *Informed* condition, to 28% in the *Not informed* condition, to 47% in the *Observed choice* condition, with a significant ($z = 2.59$, $p < .01$) difference between the extremes.

Compared to the overall sample, support for social-signaling in the subsamples is quite robust. Figure 1 plots giving frequency by condition. The conditions are arrayed in decreasing order of predicted giving (according to the social-signaling interpretation), with the conditions predicted to have equal giving pooled together. The predicted decline in giving is visible in the subsamples, but less so in the overall data. The gap between the extreme giving levels is close to 40 percentage points in both subsamples.

[14] The results for the *Observed choice* condition are the same in both subsamples because all subjects in the *Observed choice* condition of the probabilistic dictator game were assigned to the *Low Anonymity* condition of the game with variable identifiability.

Figure 1: The proportion of dictators choosing (5,5), by condition.



5 Conclusion

I present a model of a decision-maker who is concerned about how her choice affects the beliefs of an observer about her preferences. The experiment tested the hypothesis that giving is driven by concern for self-image. The test is consistent with a rational, Bayesian model and the first real-stakes test of its kind. It is based upon the theoretical insight that lowering the probability that a choice will count, while holding constant the information of an outside observer will affect the behavior of a self-signaler, but not a social-signaler.

The data do not support the hypothesis that self-image concern has a major impact on giving behavior. Though giving increases with a drop in choice probability, the effect is not statistically significant, despite a rather large sample. If self-signaling is present, it is

quite subtle and clearly overshadowed by the effects of social-signaling. On one hand, this is surprising given the wealth of psychological evidence showing the importance of self-concept. On the other hand, however, because individuals must have some degree of self-knowledge, it is plausible that an individual giving decision conveys more information to others than it does to the self, leading to stronger external effects.

A puzzle remains as to how to explain the exploitation of 'moral wiggle-room' *a la* Dana, Weber, and Kuang (2007). The abundance of suggestive evidence, but dearth of direct evidence consistent with rational model of self-signaling in giving suggests that the influence of self-image concern on giving behavior may depend upon other factors such as self-deception or the subtle influence of environmental variables on the feeling of moral obligation. Further work should investigate how giving behavior depends upon the environment in which choices are elicited, as done by Grossman (2010a).

Though the mixed support for social-signaling in the overall data is somewhat surprising, restricting attention away from potential money-maximizers yields compelling evidence that concern for the beliefs of others plays a major role in driving the giving behavior of a large subset of the population. The subjects in both subsamples analyzed above appear much more sensitive on average to the information of the recipient than the overall population. Each predicted difference across informational conditions is more than double—in both absolute and proportional terms—than the corresponding difference in the overall data.

Given the strength of previous evidence, the broad claim that givers care about how their choices impact the the beliefs of others is uncontroversial. However, the precise nature of that concern is not entirely clear. Because the model does not address the issue of the identifiability of the dictator, a social-signaler it describes could be driven by either a selfish concern for her image in the eyes any observer or a less selfish concern for how her choice affects the emotions of the recipient or his feelings about his treatment. Of these two possibilities, social-image concern has been more heavily studied and documented (e.g. Andreoni and Petrie (2004), Soetevent (2005), Alpizar, Carlsson, and Johansson-Stenman (2008)), yet the fact that the social-signaling evidence presented here (and much of the

previously cited work) is generated in laboratory experiments strict anonymity presents a challenge for this explanation. How can your choice garner esteem when no one knows that you are the one who chose it? Social-image may still be an important motivator in anonymous laboratory settings if 'rule rational' participants import externally useful behavior into the lab. Future work should distinguish between this type of behavior and disinterested, altruistic concern for the beliefs of others.

One limitation of the model is that it takes a narrow view of what people would like to signal about themselves. A person's desire to avoid coming across as concerned about her image may be as real and as powerful as her wish to be perceived as having a certain preference over outcomes. The esteem for a person who transparently gives in the most visible way possible may be undermined because her action does not appear disinterested. A more general model of social- and self-signaling might include heterogeneity in the concern for beliefs and signaling along this dimension as well.

Finally, many attributes are socially valued. A person may care about how she is perceived with respect to skill, self-control, work ethic, or racial or political attitudes and may signal to herself or to others through task choice, labor supply, affiliations and consumer choices. Inasmuch as the results of this experiment support the notion of signaling social-preferences, they lend credibility to these broader notions of social-signaling as well. Further research should explore the economic impact of the self-presentation motive across these other personality attributes.

References

Ainslie, G. (1992): *Picoeconomics*. Cambridge University Press.

Alpizar, F., F. Carlsson, and O. Johansson-Stenman (2008): "Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica," *Journal of Public Economics*, 92(5-6), 1047–1060.

Andreoni, J., and B. D. Bernheim (2009): "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica*, 77(5), 1607–1636.

Andreoni, J., and R. Petrie (2004): "Public Goods Experiments Without Confidentiality: A Glimpse Into Fund-Raising," *Journal of Public Economics*, 88(7-8), 1605 – 1623.

Ariely, D., A. Bracha, and S. Meier (2009): "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, 99(1), 544–55.

Bagwell, L. S., and B. D. Bernheim (1996): "Veblen Effects in a Theory of Conspicuous Consumption," *American Economic Review*, 86(3), 349–73.

Banks, J. S., and J. Sobel (1987): "Equilibrium Selection in Signaling Games," *Econometrica*, 5(3), 647–661.

Baumeister, R. (1998): "The Self," in *Handbook of Social Psychology*, ed. by D. Gilbert, S. Fiske, and G. Lindzey, pp. 680–740. McGraw-Hill, New York.

Bem, D. J. (1972): "Self-perception Theory," in *Advances in Experimental Social Psychology*, ed. by L. Berkowitz, vol. 6, pp. 1–62. Academic Press, New York.

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

——— (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96(5), 1652–1678.

Bernheim, B. D. (1994): "A Theory of Conformity," *American Economic Review*, 102(5), 841–877.

Bodner, R., and D. Prelec (2003): "Self-signaling and Diagnostic Utility in Everyday Decision Making," in *The Psychology of Economic Decisions, vol. Rationality and Well-being*, ed. by I. Brocas, and J. Carillo, chap. 6. Oxford University Press.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90, 166–193.

Broberg, T., T. Ellingsen, and M. Johannesson (2007): "Is Generosity Involuntary," *Economics Letters*, 94(1), 32–37.

Brown, J., and S. Smart (1991): "The self and social conduct: Linking self-representations to prosocial behavior," *Journal of Personality and Social Psychology*, 60(3), 368–375.

Burlando, R. M., and F. Guala (2004): "Heterogeneous Agents in Public Goods Experiments," *Experimental Economics*, 8, 35–54.

Camerer, C. (2003): *Behavioral Game Theory*. Princeton University Press.

Carlsmith, J., and A. Gross (1969): "Some Effects of Guilt on Compliance," *Journal of Personality and Social Psychology*, 11, 232–239.

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

Charness, G., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117(3), 817–869.

Cho, I.-K., and D. M. Kreps (1987): "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102(2), 179–221.

Dana, J., D. M. Cain, and R. Dawes (2006): "What you don't know won't hurt me: Costly (but quiet) exit in a dictator game," *Organizational Behavior and Human Decision Processes*, 100(2), 193–201.

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

DellaVigna, S., J. A. List, and U. Malmendier (2009): "Testing for Altruism and Social Pressure in Charitable Giving," Working Paper 15629, National Bureau of Economic Research.

Ellingsen, T., and M. Johannesson (2008): "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98(3), 990–1008.

Fehr, E., and K. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817–868.

Fischbacher, U. (2007): "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 10(2), 171–178.

Fischbacher, U., S. Gächter, and E. Fehr (2001): "Are People Conditionally Cooperative," *Economics Letters*, 71(3), 397–404.

Glazer, A., and K. Konrad (1996): "A signaling explanation for private charity.," *American Economic Review*, 86(4), 1019–1028.

Goffman, E. (1959): *The Presentation of Self in Everyday Life*. Doubleday, Garden City, N.Y.

Grossman, Z. (2010a): "Strategic Ignorance and the Robustness of Social Preferences," University of California at Santa Barbara, Economics Working Paper Series 1469100, Department of Economics, UC Santa Barbara.

——— (2010b): "What is Behind 'Moral Wiggle-Room'?," working paper, Department of Economics, UC Santa Barbara.

Hoffman, E., K. McCabe, K. Shachat, and V. Smith (1994): "Preferences, Property Rights, and Anonymity in Bargaining Games," *Games and Economic Behavior*, 7, 346–380.

Kunda, Z. (1990): "The Case for Motivated Reasoning," *Psychological Bulletin*, 108(3), 480–498.

Kurzban, R., and D. Hauser (2005): "Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations," *Proceedings of the National Academy of Sciences*, 102(5), 1803–1807.

Lazear, E. P., U. Malmendier, and R. A. Weber (2009): "Sorting and Social Preferences," working paper.

Lepper, M. R., D. Greene, and R. E. Nisbett (1973): "Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis," *jpsp*, 28, 129–137.

Mazar, N., O. Amir, and D. Ariely (2008): "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance," *Journal of Marketing Research*, 45(6), 633–644.

Shaw, L., C. Batson, and R. Todd (1994): "Empathy Avoidance: Forestalling Feeling for Another in Order to Escape the Motivational Consequences," *Journal of Personality and Social Psychology*, 67(5), 879–887.

Soetevent, A. R. (2005): "Anonymity in giving in a natural context–a field experiment in 30 churches," *Journal of Public Economics*, 89(11-12), 2301–2323.

Tadelis, S. (2008): "The Power of Shame and the Rationality of Trust," working paper.

A Proofs

Proof of Fact 1. Because $\sigma(\rho_1) = 0$ and $\sigma(\rho_2) = 1$, we know that

$$\begin{aligned} E[w(x,\rho_1)|1] + \lambda V(1,g) &\leq E[w(x,\rho_1)|0] + \lambda V(0,g) \qquad \text{and} \\ E[w(x,\rho_2)|1] + \lambda V(1,g) &\geq E[w(x,\rho_2)|0] + \lambda V(0,g), \end{aligned}$$

with at least one of the inequalities being strict. Rearranging them yields

$$E[w(x,\rho_2)|0] - E[w(x,\rho_2)|0] \leq \lambda(V(1,g) - V(0,g)) \leq E[w(x,\rho_1)|0] - E[w(x,\rho_1)|0],$$

or

$$C(\rho_2) \leq C(\rho_1),$$

which implies that $\rho_1 \leq \rho_2$. Because $\sigma(\rho_1) \neq \sigma(\rho_2)$, it must be the case that $\rho_1 < \rho_2$. It follows directly from this that $V(0,g) < V(1,g)$. If this were not the case then ρ_2 would have an incentive to deviate to $a_D = 0$, contradicting the assumption of equilibrium. □

Proof of Lemma 1. Evaluation and algebra. □

Proof of Proposition 1. Implicitly differentiating the equilibrium indifference condition with respect to q yields

$$\frac{\partial \rho^*}{\partial q} = \frac{-\frac{\partial C(\rho^*)}{\partial q} + \lambda \frac{\partial B(\rho^*)}{\partial q} + \lambda \frac{\partial B(\rho^*)}{\partial k}\frac{\partial k}{\partial q}}{\frac{\partial C(\rho^*)}{\partial \rho^*} - \lambda \frac{\partial B(\rho^*)}{\partial \rho^*}}.$$

In stable equilibria, the denominator must be negative (see Appendix B), so the expression takes on the opposite sign as the numerator. Monotonicity guarantees that $\frac{\partial C(\rho^*)}{\partial q} = c(\rho^*)$ is positive.

1. In the *Observed choice* condition, $\frac{\partial B(\rho^*)}{\partial q} = \lambda \frac{\partial B(\rho^*)}{\partial k} = \frac{\partial k}{\partial q} = 0$, because B is independent of q and k, and k is fixed. Thus, the numerator is negative and the overall expression is positive.

2. In the *Not informed* condition, k is fixed, so $\frac{\partial k}{\partial q} = 0$, and $\frac{\partial B(\rho^*)}{\partial q} = kB^1(\rho^*)$. The numerator is thus $-c(\rho^*) + \lambda k B^1(\rho^*) = \frac{1}{q}[B(\rho^*) - C(\rho^*)]$, which is zero in equilibrium, making the overall expression zero.

3. In the *Informed* condition, $k = q$, so $\frac{\partial k}{\partial q} = 1$, and the last term in the numerator simplifies to $\lambda q B^1(\rho^*)$, which monotonicity guarantees is positive for nonzero q. Thus the numerator can be written $\frac{1}{q}[B(\rho^*) - C(\rho^*)] + \lambda q B^1(\rho^*) = \lambda q B^1(\rho^*) \geq 0$, with equality if and only if $q = 0$.

□

Proof of Proposition 2. The proof proceeds by implicitly differentiating the indifference condition with respect to k, but first I argue that it is appropriate to use the specification of $B(\rho)$ that applies when only x is observable. When only x is observed, but $q = k = 1$, the beliefs-benefit function, $B(\rho) = qkB^1(\rho)$, is identically equal to that under *Observed choice*. Furthermore, because k is fixed at one in the *Observed choice* condition, any perturbation of k necessarily requires that either before or after the perturbation, only the outcome is observable. Thus, even if an arbitrarily small change in k was caused by a change in the observability of a_D, the resulting effect can be characterized by the partial derivative that is derived when $qkB^1(\rho)$ is substituted for $B(\rho)$, with $q = k = 1$, whenever a_D is observed.

Implicitly differentiating of the indifference condition with respect to k then yields

$$\frac{\partial \rho^*}{\partial k} = \frac{\lambda q \frac{\partial B(\rho^*)}{\partial k}}{\frac{\partial C(\rho^*)}{\partial \rho^*} - \lambda \frac{\partial B(\rho^*)}{\partial \rho^*}} = \frac{\lambda B^1(\rho^*)}{C'(\rho^*) - \lambda B'(\rho^*)}.$$

Monotonicity guarantees that the numerator is positive and in stable equilibria the denominator must be negative, therefore the expression is negative. □

B Equilibrium and the Stability Condition

Figure 2 illustrates two interior equilibria. Utility is on the vertical axis, while types are arranged on the horizontal axis *in reverse order*. The upward-sloping $C(\rho)$ curve reflects the

fact that lower types (on the right) have greater expected disutility from choosing fairly. In this example the highest types directly prefer the fair allocation, with ρ^c being indifferent. The weighted beliefs-utility benefit curve, $\lambda B(\rho)$, is strictly positive and bounded by λ, and it depends upon the cutoff.

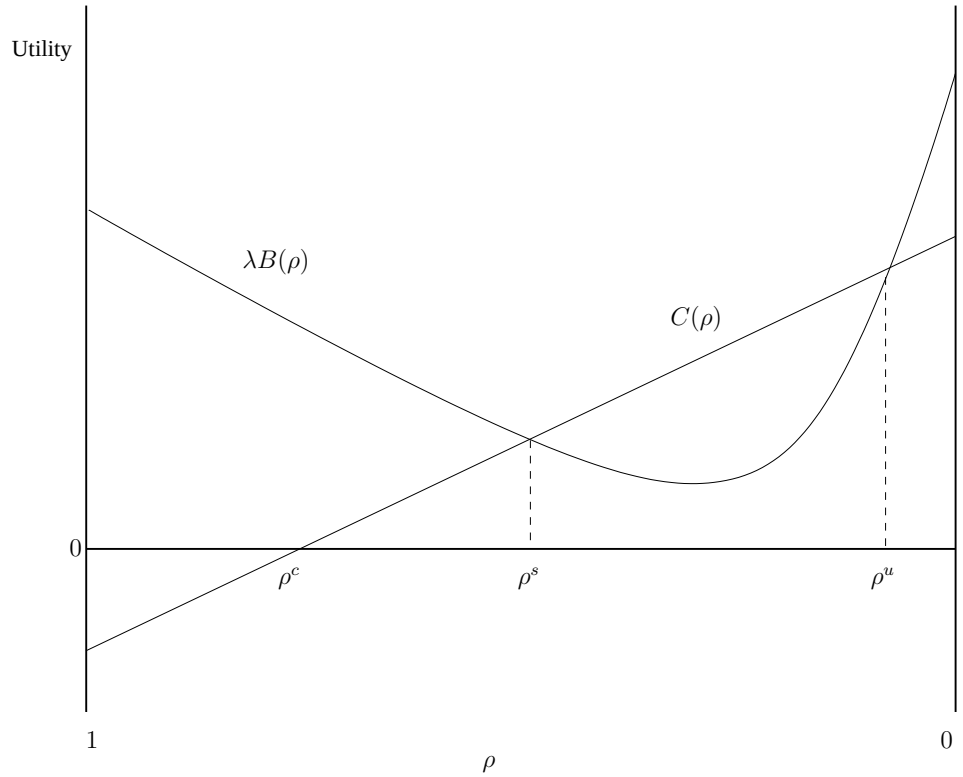


Figure 2: An illustration of stable (ρ^s) and unstable (ρ^u) interior equilibria. Note: horizontal axis has ρ going from 1 to zero.

The equilibrium cutoff occurs where the two curves intersect. Monotonicity guarantees that both the cost and benefit of choosing fairly are positive for the cutoff type, which means that the set of types between the cutoff and ρ^c disregard their strict outcome preference and choose the fair allocation for the expected beliefs-utility benefit.

The equilibrium at ρ^u is unstable. Holding constant the beliefs-utility benefit at $\lambda B(\rho^u)$, a

perturbation that results in say, a drop in $C(\rho^u)$ would provide ρ^u and some lower neighbors (to the right) the incentive to switch to $a_D = 1$. The beliefs-utility benefit increases in response, and does so more quickly than the marginal type's cost, and the effect snowballs until a different, stable equilibrium is reached. In general, an equilibrium is unstable if and only if the beliefs-utility benefit increases more quickly than the marginal type's cost when more people give, that is, when $\lambda B'(\rho) < C'(\rho)$. On the other hand, this same perturbation at ρ^s results in incentives that restore equilibrium. For this reason I restrict attention to stable equilibria.

The slope of $B(\rho)$ derives from the local curvature of the weighted value function $v_g(\rho) = v(\rho)g(\rho)$. Specifically, when $v_g''(\rho) > 0$, it is important to distinguish oneself as one of the highest types, while the difference in the valuations of lower types is not as great so $B'(\rho) < 0$. This is because when more people give, the marginal types dilute the average esteem of the givers more than they diminish the esteem of the non-givers. By the same logic, $B(\rho)$ slopes upward when $v_g''(\rho) < 0$. Thus, excluding unstable equilibria boils down to limiting the extent to which the value function can be locally concave at an interior equilibrium.[15]

C A Second Dictator Game

The second decision, used to identify potential money-maximizers, was a dictator game with variable identifiability. Participants were randomly assigned into four-person groups with each person seated in a different row. The roles of dictators 1-3, and recipient were randomly assigned within each group and each dictator was randomly assigned an endowment of 6 or 8. The recipient had no initial endowment. Each dictator independently decided how much money to transfer to the recipient and was randomly assigned an exchange rate of 1:1, 1:2, or 2:1. At the time of the decision, the endowments and exchange rates faced by the three dictators were common knowledge and after the decisions were complete all players saw a

[15] The weak restrictions imposed on the value function and on the cost curve do not preclude multiple equilibria, but existence follows from Cho and Kreps (1987), and every unstable equilibrium is accompanied by another stable equilibrium in which more types choose fairly.

Table 2: Varying identifiability in a dictator game – mean transfers

	$t > 0$		Endowment		Price		
	(%)	Overall	6	8	2	1	0.5
High Anonymity (N=69)	48	0.67	0.57	0.84	0.46	0.75	0.80
Low Anonymity (N=132)	66	1.04	0.85	1.20	1.23	0.98	0.91

summary of the endowment, exchange rate, transfer and profit of each of the three dictators and the profit of the recipient. In the *Low Anonymity* condition the recipient was also told in which row each dictator was seated, reducing the level of anonymity afforded to the dictator. In the *High Anonymity* condition the recipient was not given this information. In both conditions the informational structure of the game was common knowledge.

Though the decision is not designed to test the narrow predictions of the current model, the results, reported in Table 2, are consistent with social-signaling. The data includes the decisions of 201 dictators in 67 groups.[16] The first column displays the frequency of non-zero transfers in each condition. In the *High Anonymity* condition, 48% of the dictators transferred a non-zero amount, while 66% did in the *Low Anonymity* condition. Furthermore, in the *Low Anonymity* condition the mean transfer was 1.04, which is .37 higher than in the *High Anonymity* condition (.67).

The difference in giving frequency ($z = 2.48$, $p < .01$) and in mean transfer ($z = 2.37$, $p < .01$) are both statistically significant. The difference in mean transfers (.37) is larger than the difference in the mean transfer of dictators endowed with 8 (1.09) versus 6 (.78), suggesting that lowering anonymity by a factor of four has an effect on the recipient's payoffs similar to increasing the dictator's endowment by 2.

D Experimental Instructions

[Welcome Instructions]

[16] Not everyone in each session participated in this decision. Furthermore, one out of four participants was a recipient and does not appear in this data.

Welcome and thank you for participating in this decision-making experiment. You will be paid for participating and research foundations have provided the funds for this experiment. You will make several different decisions. Each decision is independent from each of your other decisions, so that your choices and outcomes in one decision will not affect your outcomes in any other decision. In every case, you will be anonymously paired with one or more other people, so that your decision may affect the payoffs of others, just as the decision of other people in your group may affect your payoff. Your payoff may also depend upon chance. Please pay careful attention to the instructions as a considerable amount of money is at stake.

The entire experiment should be complete within an hour. At the end of the experiment you will be paid privately and by check. Your participation in the experiment and any information about your earnings will be kept strictly confidential. Your payment-receipt, participant form and consent form are the only places in which you name or social security number are recorded. You will never be asked to reveal your identity to anyone during the course of the experiment. Neither the experimenter nor the other participants will be able to link you to any of your decisions. In order to keep you decisions private, please do not reveal your choices to any other participant.

This experiment consists of two parts. In Part 1 you will make three decisions and in Part 2 you will make two decisions. Your earnings will be calculated for each decision and at the end of experiment one decision from each part will be selected for payment. Your total earnings will be the sum of your earnings from these two decisions. During the experiment we will speak of Experimental Currency Units (ECU) instead of dollars. Your earnings will be stated in ECUs and converted to dollars before you get paid at the end of the session. One ECU is worth exactly 1.25 dollars. Thus, your earnings will effectively be increased by 25 percent when they are converted to dollars.

If you have any questions during the experiment, please raise your hand and wait for assistance. Before you proceed to Part I please note that for each screen, once you click OK you cannot go back to the previous screen. Please make sure you have read and understand

everything completely before you move on.

[Probabilistic Dictator Game Instructions: *Observed choice* condition]

INSTRUCTIONS: READ VERY CAREFULLY. IF YOU HAVE A QUESTION, PLEASE RAISE YOUR HAND AND WAIT FOR ASSISTANCE.

1. You will make two decisions, each of which can affect your payoff and the payoff of another subject. One of these two decisions will be randomly selected for payment. For each decision, you have been **randomly** and **anonymously** matched with a partner.

2. For each decision, there are two payment outcomes for you and your partner, A and B. The computer chooses either outcome A or B at random. After learning the computer's choice, you will choose one of the outcomes.

3. The computer will then select either your choice or its own choice to implement. This selection will be done randomly. The probability that the computer will select its own choice has been predetermined by the experimenter, so it will not be affected by your decision. You will learn that probability, but you will not be told whether or not your choice was selected.

4. For each decision, if your choice is selected, the profit for you and for your partner will be calculated according to the outcome you chose. If your choice is not selected, your profits will be calculated according to the outcome chosen by the computer. **Your partner will be told your choice even if it was not selected.**

5. After you have completed your decisions, your profit will be summarized for you. Your profit has two parts: one from the result of your decision, and one from the result of someone else's decision. Your total profit for each decision will be the sum of these two parts. For each decision the profit summary includes: your choice, the computer's choice, the outcome, and the profit for you and for your partner. In a separate box the other part of your profit– resulting from the decisions of others who have you as partner– will be displayed. It will include the same information: the other person's choice, the computer's choice, the outcome, and the profit for you and the person who made the decision.

[Probabilistic Dictator Game Choice Entry: *Not informed, Low* condition]

Your decisions:

Decision 1:

Outcome A: you get 7.00; your partner gets 1.00.

Outcome B you get 5.00; your partner gets 5.00.

Decision 2:

Outcome A: you get 3.75; your partner gets 7.50.

Outcome B you get 4.00; your partner gets 4.00.

1. For half of the subjects in this experiment (and in this session) the probability that the computer's choice will be implemented is zero and for the other half it is two-thirds. If the probability is zero then the outcome you choose will definitely be implemented. If the probability is two-thirds, then two-thirds of the time the computer's choice will be implemented regardless of what you choose.

2. Your probability is two thirds. This means that when you make your choice, there is a two-thirds chance that the computer's– and not your choice– will be implemented.

3. Your partner will not be informed that the computer's choice will be implemented with probability two-thirds. This means that your partner will only know that it is equally likely that you face each of the two probabilities.

Now it is time to learn the computer's choices and to make your choices.