Self-signaling and social-signaling in giving

Zachary Grossman

University of California – Santa Barbara, United States

ABSTRACT

Can a Bayesian signaling model explain patterns of giving driven by self-image concern as well as those driven by social-image? I experimentally test the predictions of such a model about how potential givers will respond to a change in the probability that their choice will be implemented. A self-signal is predicted to respond with increased giving, but he predicted response of a social-signal is increased giving, no change, or decreased giving depending on the information available to an outside observer. The experiment thus presents a test of the joint, independent and relative effects of social- and self-signaling in giving. The results provide little evidence of self-signaling, but stronger evidence of social-signaling.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

People like to be perceived favorably, by themselves and by others, but some personality attributes that carry a high social value, such as a concern for the well-being of others and fair-mindedness are not directly observable to outsiders and are difficult to introspect. However, our actions offer a window into our personality and tastes. Knowing this, people may distort their behavior to send a positive signal about themselves. Indeed, laboratory (Andreoni and Bernheim, 2009) and field (DellaVigna et al., 2012; Ariely et al., 2009) tests have shown that Bayesian signaling models, in which choices and beliefs about those making them are determined simultaneously in equilibrium, are useful for predicting and explaining patterns of giving driven by social-image concern (e.g., Lazear et al., 2012; Dana et al., 2006; Broberg et al., 2007).

* Correspondence to: Department of Economics, University of California, 2127 North Hall, Santa Barbara, CA 93106, United States. Tel: +1 805 893 7309; fax: +1 805 893 8830.
E-mail address: grossman@econ.ucsb.edu

http://dx.doi.org/10.1016/j.jebo.2015.05.008
0167-2681/© 2015 Elsevier B.V. All rights reserved.
Self-image is also an important motivator of social behavior (Bem, 1972; Baumeister, 1998) and, because one cannot perfectly introspect or recall the motivation underlying one's own behavior, a person may self-signal so as to manage her impression of herself. Economists and psychologists have argued for the role that self-signaling plays in observed patterns of giving (Murnighan et al., 2001; Dana et al., 2007; Fischbacher and Fölmi-Heusi, 2013) and in maintaining honesty (Mazar et al., 2008). If the sender and receiver roles are viewed as two aspects of a divided self, rather than separate people, Bayesian signaling models can also have a self-signaling interpretation and can be used to explain or predict behavior. In this interpretation, even though one can act upon one's preferences, the part of the self acting as the observer does not have direct access to them and instead must infer them from actions, just like an outside observer would.1

However, despite the successful track record of the Bayesian-signaling-game approach to modeling social–image concern, direct evidence of Bayesian self-signaling remains elusive. The evidence from psychology depicts altruistic acts as a consequence of self-image management, but not as the instrument.2 Partly behind the lack of direct tests of Bayesian self-signaling is the fact that, while it is easy to manipulate the information an external observer has about a person’s choice so as to document the resulting change in behavior, one cannot manipulate in the same manner the observability of one’s choice to oneself. As a consequence, the existing experimental evidence of self-signaling, discussed in more detail below, is indirect and suggestive, lacking a direct test of the predictions of Bayesian signaling equilibrium.

I conduct a direct experimental test of the hypothesis that a Bayesian signaling model can predict and explain patterns of giving driven by self-image concern, with complementary tests of social–image concern. Participants play a binary dictator game with, in two separate conditions, High or Low probability that their choice will actually count, instead of being overridden by the computer. Lowering this probability shrinks the expected cost (whether positive or negative) of the giving gesture, but social-preference models based exclusively on outcomes predict that this will not affect the dictator’s choice. In contrast, in a signaling model of image concern, presented in Section 2, the same probability change can have an impact because the dictator balances the expected material cost of pledging to give against the expected image benefit.

When the dictator’s audience can directly observe her action, lowering the probability that it counts affects neither the dictator’s ability to influence the signal nor the observer’s perception of the quality of the signal. The only impact is to cheapen the outcome–cost of the giving gesture, leading to a higher giving rate in equilibrium. However, if only the outcome is observed, the same change also reduces the dictator’s ability to influence the signal she sends, resulting in an unchanged or, if the probability drop is also observable and thus undermines the informational content of the signal, a diminished giving rate.

The experiment, described in Section 3, features recipients who have been placed in one of the three information environments described above, in which the dictator’s Choice is directly observed, only the realized Outcome is observed, or both the Probability & Outcome are observed. Importantly, the model’s predictions about behavior in the experiment hinge on whether the decision-maker is viewed as a social-signaler or a self-signaler. In the case of the former, the predicted response adheres to the aforementioned pattern: increased giving, no change, and decreased giving, respectively.

However, even if the dictator is interpreted as a self-signaler who cannot directly observe her own preferences, she always can directly observe her own actions. So even as I manipulate the recipient’s information in the experiment, the self-signaling interpretation of the model regards the dictator as always being in the information condition in which her Choice is observable. While the social-signaler’s predicted response to a drop in probability varies by experimental information condition, the self-signaler is predicted to be more likely to give across all three. This distinction between the two interpretations of the model is what creates the ability to identify Bayesian self-signaling in giving independently of social signaling, which is the main empirical contribution of this paper.

The test consists of comparing the rate at which dictators commit to giving under High versus Low probability in the Outcome condition. Social-signalers are not predicted to change their behavior at all in response to the probability change, but self-signalers are predicted to respond with a higher giving rate. While the rate of giving does increase slightly, from 20% to 25%, this increase is not statistically significant, despite a very large sample. Furthermore, in the Probability & Outcome condition the two forms of signaling are predicted to act in opposition. In this case, the frequency of giving falls significantly, in line with social-signaling and contra self-signaling. Thus, the results do not offer conclusive support for the usefulness of Bayesian signaling models to predict patterns of giving driven by self-image concern.

A second contribution of the paper is a vigorous test of Bayesian social-signaling, which predicts a weak hierarchy of giving rates across the six cells of the 2 × 3 design. While some experimental conditions feature giving rates that conform to the predicted relationships, the data are relatively noisy and one cannot reject the hypothesis that the giving rate is the same across all conditions. That said, in a large subsample chosen to exclude potential “money-maximizers”, social-signaling is quite pronounced. Behavior in this sample closely follows the predicted pattern, featuring swings in the giving rate of over 35 percentage-points.

---

1 Bodner and Prelec (2003) introduce the dual-self signaling model approach adopted by others such as Benabou and Tirole (2006) and used herein. 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.
2 Carlsimth and Gross (1969) find that feeling guilty about recent harmful behavior can lead subjects to be more compliant with requests for help. 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. Shaw et al. (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.
Because the support for social signaling is so mixed, I conducted a follow-up experiment, described in Section 4, designed to address features of the main experiment that may have added noise. Participants made only one decision, instead of several, and they gave anonymously to a large charity, instead of to a fellow participant. The results of the follow-up experiment reinforce the conclusions of the main experiment, offering little support for self-signaling and bolstering the support for social-signaling.

In summary, the influence of the social-image motive on giving is quite clear and is largely consistent with signaling equilibrium, but the influence of self-image concern on giving is less evident from these results. To the extent that self-image does have the impact on giving suggested by previous studies, it may involve reasoning and cognitive processes not consistent with Bayesian signaling equilibrium or it may influence giving through indirect channels, such as the acquisition of information related to a giving decision.

This paper adds to a growing body of evidence on self-signaling and the motivation for giving. Previous experimental evidence has suggested self-image as a motive for giving, typically by eliminating alternative explanations for an observed phenomenon that is consistent with some form of self-signaling, such as the regularity of willful ignorance of how one's choice affects another (Dana et al., 2007) or the pattern of self-reported, incentivized die rolls (Fischbacher and Föllmi-Heusi, 2013), Cueva and Dessi (2012) and Ploner and Regner (2013) find increased giving following manipulations that increase the salience of self-image: letting subjects observe public statements of donations of others and having behaved anti-socially in an earlier stage of the experiment, respectively. Tonin and Vlassopoulos (2013) find that subjects opt out of a probabilistic decision to give ex post and argue that this reflects a reduced signaling benefit stemming from the satiation of self-image motivation, as well as the increased cost of giving ex post.

While this evidence may be consistent with some form of self-signaling, it does not provide a direct test of the predictions of Bayesian signaling equilibrium. The closest and most direct evidence of self-signaling in giving comes from Grossman and van der Weele (2013), who find evidence of self-signaling in social decisions. However, this evidence is mainly seen in information acquisition and not directly in a simple giving choice. Looking at honesty, as opposed to giving, Mazar et al. (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. Thus, individuals may distort their behavior so as to maintain their self-concept, but this self-image management depends upon self-deception and non-Bayesian rationalizations. This kind of self-deceptive image management is certainly consistent with the preponderance of evidence suggesting an influence of self-image on giving. Nonetheless, the current experiment does not offer conclusive evidence in support of the predictions of Bayesian self-signaling.

2. A signaling model of a probabilistic dictator choice

In this section, I present a model of a decision-maker (D, female pronouns) playing a binary dictator game and derive the comparative-static predictions upon which the experiment is based.\(^3\) As in Andreoni and Bernheim (2009), the dictator-game outcome is probabilistic, meaning that with some probability the outcome of the game will be determined by chance instead of her choice.\(^4\) The decision-maker has preferences over the outcome of the game and also cares about how her choice affects the endogenously determined beliefs of a passive observer (O, male pronouns) about those preferences. O cannot directly observe D’s preference type, but does observe a signal generated by D’s choice and uses it to update her prior beliefs. Knowing this, D trades off the instrumental utility from the outcome resulting from her action with the image utility derived from the signal it generates. 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.

Like 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. The key assumption that is crucial for separately identifying self- and social-signaling in the experiment is that a person can always observe her own action. This means that for a self-signaler, O always receives a perfectly informative signal of D’s choice, while the same is not true for social-signalers.

2.1. Timing and information

First nature draws D’s preference type, \(\rho\), from a continuous distribution with full support over the unit interval, and \(q > 0\), the probability that D’s choice will count. D observes \(q\) and, in some information conditions, O does as well. Then D and nature simultaneously choose \((a_d \text{ and } a_n, \text{ respectively})\) from the set \(\{0, 1\}\), with 1 corresponding to the more fair or generous option in the dictator game.\(^5\) The outcome of the game, \(a\), is then determined. With probability \(q\), the dictator’s

---

\(^3\) Although the binary choice stands in contrast to the continuous models of Benabou and Tirole (2006) and Andreoni and Bernheim (2009), it simplifies the comparative-static analysis and the experiment.

\(^4\) The model may be applied to any situation in which a person decides whether or not to commit a costly act that stochastically affects the outcome.

\(^5\) Neither the process by which nature’s choice is determined, nor the realization of the choice, is relevant for the analysis of the dictator’s equilibrium behavior, as characterized by the two propositions below. One may assume that there is some commonly known distribution from which \(a_n\) is chosen. This also means that any experimental tests based upon equilibrium predictions are not predicted to be sensitive to design choices regarding nature’s realization.
choice is implemented as the outcome \((a = a_d)\) and with probability \(1 - q\), nature’s choice is implemented \((a = a_n)\). Finally, \(O\) observes a signal \(\tilde{a}\), described below, and updates his beliefs about \(\rho\).

The distributions from which \(\rho, q,\) and \(a_n\) are drawn are common knowledge and the comparative-static predictions are derived under three separate information conditions for \(O\):

- D’s Choice is observed \((\tilde{a} = a_d)\);
- the Probability & Outcome are both observed \((O\) observes \(q\) and \(\tilde{a} = a)\);
- only the Outcome is observed \((O\) knows only \(E[q]\) and \(\tilde{a} = a)\).

2.2. Preferences

D’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. Thus, \(\rho\) 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). While \(c(\rho)\) may be positive or negative, the crucial sorting assumption is that \(c'(\rho) < 0\) and thus, \(C(\rho) < 0\), implying that higher types find it less costly to choose fairly.

The decision-maker also cares directly about O’s beliefs about her type. Because \(\rho\) is private, the decision-maker cares about the expectation of \(\rho\) taken over O’s beliefs, which are updated after observing \(\tilde{a}\). An interpretation function, \(g\), maps O’s signal to posterior beliefs, with \(g(\rho, \tilde{a})\) denoting the updated probability that D has type \(\rho\), conditional on observing \(\tilde{a}\). 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 \(V(a_d, g) = E[\int rg(\rho, \tilde{a})d\rho | a_d]\), where expectations are taken over the possible realizations of \(\tilde{a}\) 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, \(\lambda\), is zero, the model reduces to one of purely outcome-based social preferences. In this degenerate case, all of the comparative statics derived below predict zero effects and no purely outcome-based model would generate any of the results detailed below.

2.3. Equilibrium

The solution concept that I apply is a form of Perfect Bayesian Equilibrium for psychological games. Actions and beliefs are determined simultaneously. All types must maximize utility, taking the observer’s interpretation as given, and the observer’s interpretation must 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 \(q\). Equilibrium is defined over a function, \(\sigma\), which assigns an action to each type, and an interpretation function, \(g\), which maps observations to posterior beliefs. Equilibrium requires that

- for each type \(\rho\), \(U(\sigma(\rho), \rho, g)\geq U(\sigma'\rho, \rho, g)\) for any \(\sigma'\) and
- \(g\) follows Bayes rule, when appropriate.

I refine equilibrium in two ways. First, I eliminate equilibria that are unstable in the sense that small deviations from equilibrium induce disequilibrium incentives that drive behavior further from equilibrium. Second, I require inferences following out-of-equilibrium choices to place probability only on those types with the strongest incentive to deviate. This refinement, akin to D1 (Cho and Kreps, 1987; Banks and Sobel, 1987), eliminates pooling equilibria based upon unreasonable beliefs and is used by Bernheim (1994) and Dufwenberg and Lundholm (2001). The sorting assumption implies that at most one type is indifferent in equilibrium, so, without loss of generality, I further restrict attention to pure strategies. This also guarantees that equilibrium is monotonic, in the sense that the fair allocation is chosen by higher types and therefore confers more favorable beliefs.

For non-pooling equilibria, this monotonicity allows one to restrict attention to strategies and beliefs that can be characterized by a cutoff \(\rho^*\), above which all types choose to give and below which all types choose not to give. Let \(V(a_d, \rho)\)

\[\text{Theoretical underpinning for dynamic psychological games, of which this model is an example, and for the solution concept is provided by Battigalli and Dufwenberg (2009).}\]
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 \(\rho\). 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.\(^7\)

### 2.4. Comparative Statics

Next I derive predictions about the effect on the equilibrium cutoff, \(\rho'\), of an exogenous change in the dictator’s choice probability, \(q\). Begin by observing that changing \(q\) not only impacts D’s ability to influence the outcome; it also affects her ability to influence O’s signal and the quality of that signal. Let \(k\) denote the ex ante probability that \(\hat{a}\) will take its value from \(a_d\). In other words, \(k\) represents the quality of \(\hat{a}\) as a signal of D’s choice, from O’s perspective. As a rule, \(k = 1\) in the Choice condition, while \(k = q\) in the Probability & Outcome condition, and \(k = E[q]\) in the Outcome condition. The effect on \(k\) of a change in \(q\) depends on the information condition. In the Probability & Outcome condition, \(k\) changes in lockstep with \(q\), but in the other conditions the value of \(k\) is independent of \(q\).

Let \(B^l(\rho)\) 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, when O’s signal is instead the outcome (\(\hat{a} = a\)) the beliefs-utility benefit of choosing fairly is proportional to the likelihood that D’s choice will actually influence O’s signal and to the quality of O’s signal. Specifically, \(B(\rho) = qkB^l(\rho)\).

**Proposition 1** establishes how the behavioral effect of changing the realized probability that D’s choice counts varies sharply across the three information conditions.\(^8\)

**Proposition 1.**

1. \(\frac{\partial \rho'}{\partial q} > 0\) in the Choice condition;
2. \(\frac{\partial \rho'}{\partial q} = 0\) in the Outcome condition;
3. \(\frac{\partial \rho'}{\partial q} \leq 0\) in the Probability & Outcome condition, with equality if and only if \(q = 0\).

Regardless of O’s information, lowering D’s choice probability 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, as is the case in the Choice condition. However, in the Outcome condition, the same drop in \(q\) also reduces D’s ability to influence the O’s signal. Because the outcome and signal are one and the same, the reduction in the beliefs-utility benefit of choosing fairly exactly matches the change in the outcome-utility cost, leading to zero net effect. In the Probability & Outcome condition, lowering \(q\) also leads O to further discount the quality of the signal. This further lowers the expected beliefs-utility benefit, resulting in less than the original amount of giving.\(^9\)

**Proposition 2** characterizes the qualitative effect of a change in perceived signal quality independent of \(q\), effectively yielding predictions about the effect of changing information conditions, while holding D’s choice probability constant.

**Proposition 2.** \(\frac{\partial \rho'}{\partial k} < 0\)

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. More informative signals will have a greater impact on posterior beliefs, increasing the beliefs-utility incentive to choose fairly and lowering the equilibrium cutoff.\(^10\) Note that neither proposition depends upon the identity of the observer, just the dictator’s choice probability and the observer’s information.

---

\(^7\) Equilibrium need not be unique, but a sufficient condition for uniqueness and stability is that \(\lambda B(\rho) \geq C\), which implies that the distribution from which types are drawn not increase too steeply anywhere on its domain (see also Benabou and Tirole (2006)).

\(^8\) All proofs are presented in Appendix A.

\(^9\) 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 et al., 1973). However, unlike the in model of Benabou and Tirole (2006), motivational crowding out can occur if not 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. Multi-dimensional preferences might be a reasonable assumption when 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 \(q\) is incompatible with a distributions-based model.

\(^10\) This formalizes the observation of Dana et al. (2007) that obscuring the link between actions and their consequences (in terms of beliefs) diminishes fair behavior.
3. Experiment 1

The main experiment consisted of twenty-four experimental sessions, each with 8–24 participants, conducted at the Experimental Social Science Laboratory (XLab) at the University of California-Berkeley. 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.

3.1. Design

Participants 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 provided in the Appendix. 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.

Participants played a probabilistic dictator game with structure and timing as described in Section 2. 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 (3,5), 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 and the dictator was informed of this choice in advance. For half of the dictator-recipient pairs it was (7,1) and for the other half it was (5,5). The computer also realized the uncertainty as to whether the dictator’s choice would count.

The experimental conditions followed the 2 × 3 design described in Section 2. 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 and each recipient was assigned either the Choice (C), Outcome (O), or Probability & Outcome (PO) information condition. 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. Furthermore, dictators were made aware of the information available to their recipient match.

Applying the model to the experiment, I make two crucial assumptions. First, I assume (5,5) to be the fair outcome. The payoffs were chosen precisely because the (5,5) outcome is more equitable, features a higher minimum payoff, and greater total payoff than (7,1), so any standard distribution-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. 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 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 Choice condition regardless of the experimental information condition.

For each proposition, each interpretation of the model generates a separate prediction about the proportion of dictators choosing the fair allocation. For self-signaling, Proposition 1 predicts that in every experimental information condition the giving rate will be higher when the probability is Low than when it is High, summarized as CH < CL, OH < OL, and PH < PL. In contrast, the response of social-signaleurs is predicted to vary with the information: CH < CL, OH = OL, and PH > PL. Thus, the three respective information conditions offer insight into the joint, relative, and independent effects due to self- and social-signaling. In Choice the effects work in tandem, in Probability & Outcome they are opposed, and in Outcome only self-signaling predicts an effect, thereby identifying the effect of self-signaling independent of social-signaling.

Because the maintained assumption of the self-signaling interpretation of the model is that the Observer has the same information across all experimental conditions, Proposition 2 carries no implications for the behavior of a self-signaler. Specifically, it predicts that for a given probability, the giving rate will be constant across (experimental) information conditions. In contrast, the social-signaler’s behavior is predicted to vary with the information condition. In the PH condition, k = q = 1 and thus it is structurally equivalent to CH, with behavior predicted to be the same. The recipient’s signal is more noisy in the OH condition, however, so less giving is expected. Similarly, for Low probability dictators, giving is predicted to fall as k decreases from the CL to the OL to the PL condition. Under social-signaling, the two propositions thus predict a weak hierarchy of giving across the six experimental conditions:

\[ PL < OL = OH < PH = CH < CL. \]

\[ 11 \] In addition to the probabilistic dictator game reported, participants also played one in which the payoffs were (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 (efficiency, 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.

\[ 12 \] The computer chose each outcome with equal probability and both players were informed of this choice.

\[ 13 \] The dictator knew of both possible probabilities before being told which she would face.
3.2. Results

Of the 379 participants in the main experiment, 97 (26%) chose (5,5). Table 1 summarizes the frequency of choosing (5,5) by condition, with the first two columns presenting the full data from Experiment 1. After first considering the self-signaling test, I then examine the social-signaling predictions.

Among dictators in the Outcome condition, 20% chose (5,5) in the High condition, compared to 24% in the Low condition. Though the direction of change is as predicted for self-signalers, the difference in proportions is not statistically significant ($z = 0.61$, $p < 0.27$)\(^{14}\) so the direct test offers little support for self-signaling. With a sample of almost 200 participants, the test is quite powerful. Any proportion over 0.31 in the Low condition would have been sufficient to reject at the 5% level the null hypothesis of no self-signaling. Furthermore, giving drops from the High to Low conditions, both in the Choice condition (35–25%) and Probability & Outcome (36–20%), which in both cases is inconsistent with self-signaling.

Turning to the social-signaling predictions, the support from the overall data is mixed. Recall that of six experimental conditions, the social-signaling model predicts the same level of giving in two pairs of conditions (CH-PH and OH-OL), yielding four predicted levels of giving. Fig. 1 plots giving rates by predicted giving level, with the conditions arrayed in decreasing order of predicted giving rates and with the conditions predicted to have equal giving pooled together. The pattern of the giving rates in the full sample, indicated by the green line, could either be perceived as a general downward trend with a low outlier in the CL condition or as a flat trend interrupted by high outliers in the CH and PH conditions.

Statistical analysis paints a similarly mixed picture. While the drop in Choice is not consistent with social-signaling, the difference is not statistically significant ($z = 0.98$, $p < 0.17$). On the other hand, in the Probability & Outcome condition the drop is consistent with social-signaling, and the difference is significant ($z = 1.97$, $p < 0.03$). As predicted, when the choice probability is High, the rate of fair behavior is virtually identical in the Choice (35%) and Probability & Outcome (36%) conditions, and exceeds that in the Outcome condition by a significant margin, whether considered separately ($z = 1.80$, $p < 0.04$ for CH vs. OH; $z = 2.14$, $p < 0.02$ for PH vs. OH) or pooled ($z = 2.34$, $p < 0.01$). However, among the dictators in Low, the percent giving in the Probability & Outcome, Outcome, and Choice conditions is 20, 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.

---

\(^{14}\) Hypothesis tests are one-tailed $z$-tests of proportions with pooled samples, unless otherwise noted.
The most appropriate test of the social-signaling predictions is of the joint hypothesis that the giving rates across all conditions conform to the predicted weak hierarchy. A chi-square test cannot reject the hypothesis ($\chi^2(5) = 7.82, p = 0.17$) that the giving rate is the same in all conditions. Furthermore, an analogue of Jonckheere’s trend test for dichotomous data (Neuhäuser and Hothorn, 1998) does not reject the hypothesis that the giving rate is the same in all conditions against the alternative hypothesis of the predicted trend ($T = 0.07, p = 0.33$).\(^{15}\)

This lack of variation is consistent with the hypothesis that behavior is primarily driven by distributional concerns. In light of previous evidence rejecting 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 bundling multiple decisions in the same session, doubling up roles, and complicated, computer-based instructions. The next two steps of the analysis offer attempts to assess more definitely the validity of the self- and social-signaling predictions, first by weeding potential money-maximizers out of the overall data and then by replicating the experiment in a different environment.

### 3.3. Analysis of a subsample excluding money-maximizers

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 15–35%) 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 et al., 2001; Kurzban and Hauser, 2005; Burlando and Guala, 2004).

Given the presence of such selfish types, the model will not accurately predict the behavior of all participants. Indeed, the analysis described above offers only mixed support for the model’s predictions. Many giving environments, however, might plausibly lack such selfish types. For example, a not-for-profit organization seeking to understand how to improve its fundraising outcomes can plausibly assume that its existing donor list excludes purely selfish types. Knowing whether unselfish types have a purely distributional concern for the wellbeing of others or they are subject to image concerns would be useful for their efforts to improve fundraising and would speak broadly to the usefulness of signaling models for understanding giving behavior in general. Might the signaling model still be useful to explain the behavior of the sizeable subset of the population who are not money-maximizers? To investigate the extent to which signaling effects exists within this sizeable subset of the population, I repeat the data analysis on a subsample that excludes participants categorized as likely money-maximizers.

To identify the participants most likely to be selfish types, I use 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-signalizer. 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 tests the model’s predictions.

Briefly, in this game dictators chose an amount, $t > 0$, 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.

Two hundred and one of the 379 participants played the dictator game with variable anonymity, 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, who constitute 22.3% of the subjects who participated in both dictator games, were unswayed to give by concern for the payoffs of others, self-image, and—with limited anonymity—the beliefs and potential responses of others. Define a subsample of the overall subject population by excluding these 45 participants, thereby reducing the proportion of money-maximizing subjects.

I now examine results of the probabilistic dictator game for the 156 participants that remain in this subsample, which are summarized in the third and fourth columns of Table 1. The most striking feature is the increased variation in giving rates across conditions, relative to the overall data, with the values ranging from over 45% in the *CL* and *PH* conditions to 11% in the *PL* 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 ($\chi^2(5) = 12.94, p < 0.024$). Furthermore, the Neuhäuser–Hothorn test rejects the same null hypothesis in favor of the alternative that the giving rate follows the weak hierarchy predicted by the social-signaling model ($T = 0.55, p = 0.02$).

---

15 While the Neuhäuser–Hothorn test is exact, the size of my sample renders the computation of the test-statistic’s distribution computationally impractical. Instead, I derived the distribution upon which the p-value is based using 10000 Monte Carlo simulations of the realization of the data under the null hypothesis that the probability of a subject choosing to give is 97/379 in every condition.
While the absolute increase in giving rates observed in the self-signaling test is greater than in the overall data, the evidence of self-signaling remains weak. Lowering the dictator’s choice probability in the Outcome condition increases the percent choosing (5,5) from 17 to 28, but the difference is not statistically significant (z=1.01, p < 0.16). However, the failure to reject the hypothesis that the giving rates are equal is consistent with the social-signaling prediction generated by Proposition 1, as is the drop in giving from 48% to 11% (z=2.88, p < 0.01) in the Probability & Outcome condition.

Indeed, the giving pattern in this subsample is much more consistent with social-signaling than in the overall data, suggesting that the excluded participants contributed disproportionately to the noise and that the signaling effects persist in a large fraction of the population. When the dictator faces High probability, the giving rates in the Choice (0.33) and Probability & Outcome (0.48) conditions are not significantly different (z=0.98, p < 0.33), and the 0.18 in the Outcome condition is significantly different from Probability & Outcome condition (z=2.44, p < 0.01) and the giving rate of pooled Choice and Probability & Outcome conditions (z=2.21, p < 0.02). For Low dictators, giving increases monotonically from 11% in the Probability & Outcome condition, to 28% in Outcome, to 47% in Choice, with a significant (z=2.59, p < 0.01) difference between the extremes.

4. Experiment 2
4.1. Design

The goal of Experiment 2 was to provide more conclusive results from the hypothesis tests, both by generating more data and by addressing factors that may have contributed noise to the data from Experiment 1. It featured the same basic structure as the main experiment but varied some aspects of the implementation. Instead of giving money to other participants in the same sessions, dictators gave to a large charity, for which the contribution was a small drop in an ocean of donations and which dispersed the funds across millions of recipients who will never know where the money originated. With standard measures maintaining privacy between subjects, this ensured that the only possible external audience for the dictator’s decision was the experimenter.

Furthermore, these sessions were short, lasting 15–20 min, and required participants only to make a single allocation decision. Instead of a computerized interface with silently read instructions, instructions were read out loud by the experimenter and participants made decisions by choosing an envelope. These measures were taken so as to reduce the confusion about the instructions, distractions from the task at hand, and possible portfolio effects.

The experiment consisted of fourteen experimental sessions, each with 7–18 participants. They were conducted at the Experimental and Behavioral Economics Laboratory (EBEL) at the University of California, Santa Barbara. The 185 participants were drawn from the EBEL subject pool, largely comprised of UCSB students, using the online system ORSEE (Greiner, 2003). At the beginning of the session, participants were given some basic information describing the American Cancer Society (ACS). Then they chose whether to keep $5 and donate $1 to the ACS (5,1) or keep $1 and donate $10 to the ACS (1,10), with the understanding that with some probability their choice would not count and that the (5,1) allocation would instead be implemented, regardless of the participant’s choice. In High probability sessions q = 5/6, while q = 1/6 in sessions in the Low condition. Whether the choice counted was implemented by a die rolled by a laboratory assistant.

At participants’ desks were two sealed full-view window envelopes which represented the two allocation choices. Each contained two smaller full-view window envelopes, a white one intended for the participant and displaying a single bill of the appropriate denomination (5 or 1 USD, respectively) and a purple one intended for the ACS and displaying 1 or 10 USD, respectively. The (1,10) envelope featured a lower payment for the dictator, a larger donation, and a larger surplus, so it may reasonably be regarded as being associated with more positive posterior beliefs.

To make their decisions private, participants kept the envelope they chose face-down on the desk and discarded the other one by placing it face-down through a slit in a covered discard box. One at a time, the participants took their chosen envelope to a desk, keeping it face down for privacy, where an assistant sat seated with a stack of extra (5,1) envelopes, another discard box, and a die. The assistant rolled the die which determined whether or not the participant’s choice counted. If it did not count, the participant placed her chosen envelope face-down through the slit in the covered discard box, took one of the surplus (5,1) envelopes, and proceeded to the exit. If the choice did count, the participant kept her original envelope, but still placed an envelope in the discard box, using one from the stack of extra (5,1) envelopes. This maintained similarity between the actions taken by those whose choice did count and those whose choice did not, further guarding privacy. Before exiting, the participant opened the large envelope, retained the white payment envelope, signed a receipt and left it and the purple ACS envelope in a box by the door. The experimenter was able to monitor that participants followed these instructions by observing the envelope colors, but did not observe the envelope contents unless called for by the information condition protocol.

Regardless of the information condition, recording the contents of the first discard box at the end of the session revealed the envelopes participants originally discarded, and thus, their aggregate giving rate. The information conditions were implemented as follows:

- **Choice:** After making their initial choices, participants lifted the chosen envelope up off the desk so that the experimenter and only the experimenter was able to see the envelope’s contents, and thus the choice.
• **Probability & Outcome**: The initial choice remained private, but as the participant exited the experimenter observed the resulting allocation, which was noted in the instructions as a step taken to verify that the donation amount matched the amount on the receipt. The probability condition was common knowledge as it was stated openly in the instructions.

• **Outcome**: This condition featured a minor departure from the $2 \times 3$ design of Experiment 1. Instead of maintaining the experimenter as the external audience who observes the outcome but not probability, the self-signaling test was implemented without *any* external audience. Participants’ choices were completely private—no other person ever observed either the choice or the outcome. Only session-level aggregate choices were observed. While this maintained the validity of the self-signaling test, the lack of external audience renders this condition not relevant to the social signaling test and it is excluded from the social-signaling hypothesis tests.

### 4.2. Results

The results of Experiment 2 reinforce those of the main experiment, with the main difference being a much higher overall giving rate of 50%. The *Outcome* condition offers little support for self-signaling, with 52% giving when the probability is *High* and 55% giving when the probability is *Low*, a difference that is not statistically significant ($z = 0.29, p < 0.39$). While the 47% giving rate in the PH condition is not higher than the 52% in the OH condition, as predicted by the social-signaling interpretation of the model, the pattern in the rest of the conditions is consistent with the predicted weak hierarchy of giving rates, particularly when the conditions not predicted to be different, CH and PH, are pooled together. Though the differences in giving rates are relatively small and not all large enough to achieve statistical significance, the conditions at the two extremes—CL at 68% and PL at 42%—are different enough to be significant ($z = 1.93, p < 0.03$). Furthermore, the Neuhäuser-Hothorn test rejects the null hypothesis that the giving rate is the same in all conditions in favor of the alternative hypothesis of the trend predicted by the social-signaling model ($T = 0.41, p = 0.03$).

Despite the noise in the overall data, the broader picture painted by the two experiments as well as the subsample analysis in Experiment 1 is one of general support for social-signaling predictions and little support for the self-signaling predictions. Returning to Fig. 1, the giving rate in the CL condition of the main experiment is too low relative to the other conditions to be consistent with the social-signaling predictions, but this case proves to be the exception to the broader pattern of decline that is readily visible in the figure. The subsample excluding money-maximizers, Experiment 2, and the pooled data from both experiments all exhibit the predicted decline in giving rates.

### 5. Conclusion

I conducted two experiments to test the self-signaling and social-signaling predictions of a Bayesian signaling model, exploiting the fact that the information environment for self-signals remains constant even as I manipulate the information available to outside observers. The results do not support the Bayesian self-signaling predictions. Though giving increases with a drop in choice probability, the effect is not statistically significant, despite a rather large sample.

If Bayesian 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, 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. Furthermore, signals to an internal audience in a laboratory experiment have many substitutes outside of the lab, such as giving one’s earnings to charity or to a different recipient. In light of this, it may be premature to draw strong conclusions against the self-signaling model.

In contrast, the data are largely consistent, despite some noise, with the social-signaling predictions. The subsample analysis shows that the model is quite useful in explaining the behavior of a large and identifiable part of the population, even when the full sample is noisy. Experiment 2 reinforces the conclusions of the first experiment and provides more solid evidence of social-signaling, consistent with the model’s very strong predictions.

To the extent that other evidence suggests that self-image concern has an important role shaping behavior, the lack of evidence of Bayesian self-signaling challenges the notion that self-signaling and social-signaling can be treated as two manifestations of the same phenomenon, namely distorting the behavior so as to alter the flow of information to a rationally-infering observer. While much social-signaling behavior can be explained and predicted by a rational signaling model with Bayesian inference, a successful self-signaling strategy, such as the exploitation of ‘moral wiggle-room’ *a la* Dana et al. (2007), seems to manifest primarily through information acquisition choices (Grossman and van der Weele, 2013). Alternatively, it involves an extra degree of self-deception, imperfect inference, or the subtle influence of environmental variables on the feeling of moral obligation (Grossman, 2014).

One can apply the theoretical insights from Section 2 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-signaler. However, because the dictator’s choice probability is public information, but her choice is private, the experimental manipulation parallels that of the *Probability & Outcome* 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.

Although the self-signaling test did not provide evidence of Bayesian self-signaling, it is natural to wonder whether there other ways to construct such a test, for example, in a deterministic choice environment. Having a dictator choose between...
an unfair and more fair sharing proportion, but varying the size of the pie being divided is another way to manipulate the cost of the giving gesture. Might conducting this exercise in an environment that holds constant the recipient’s information constitute another self-signaling test? Unfortunately, this particular manipulation might be expected to affect the behavior of individuals with exclusively outcome-based preferences (unless unreasonable restrictions were imposed on those preferences), undermining its validity as a self-signaling test. Another test for Bayesian self-signaling would have to take a different form.

Given the strength of previous evidence, the broad claim that givers care about how their choices impact 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 et al., 2008), yet the fact that the social-signaling evidence presented here (and much of the previously cited work) is generated in laboratory experiments with strict anonymity presents a challenge for this explanation. How can your choice garner social 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.

One limitation of the model is that it takes a narrow view of what people would like to signal about themselves. First, a person may respond to both self-image and social-concern. Indeed, the possibility of non-separable interaction between self- and social-signaling may be behind the apparent noise in the overall data. Second, 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 credence 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.

Appendix A. Proofs

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

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

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 Choice condition, $\frac{\partial B(\rho^*)}{\partial q} = \lambda \frac{\partial B(\rho^*)}{\partial k} = \frac{\partial b(\rho^*)}{\partial k} = 0$, 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 Outcome condition, $k$ is fixed, so $\frac{\partial b(\rho^*)}{\partial q} = 0$, and $\frac{\partial B(\rho^*)}{\partial q} = k B^1(\rho^*)$. The numerator is thus $-C(\rho^*) + \lambda k B^1(\rho^*) = \frac{1}{\lambda} [B(\rho^*) - C(\rho^*)]$, zero in equilibrium, making the overall expression zero.

3. In the Probability & Outcome condition, $k=q$, so $\frac{\partial b(\rho^*)}{\partial q} = 1$, and the last term in the numerator simplifies to $\lambda q B^1(\rho^*)$, which is always positive (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 benefits-benefit function, $B(\rho) = qk B^1(\rho)$, is identically equal to that under Choice. Furthermore, because $k$ is fixed at one in the 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 $d_0$, the resulting effect can be characterized by the partial derivative that is derived when $qk B^1(\rho)$ is substituted for $B(\rho)$, with $q=k=1$, whenever $d_0$ is observed.
Implicitly differentiating the indifference condition with respect to $k$ then yields

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

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

**Appendix B. Equilibrium and the stability condition**

Fig. B.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 $\rho^*$ being indifferent. The weighted beliefs-utility benefit curve, $\lambda B(\rho)$, is strictly positive and bounded by $\lambda$, and it depends upon the cutoff.

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 $\rho^*$ disregard their strict outcome preference and choose the fair allocation for the expected beliefs-utility benefit.

The equilibrium at $\rho^*$ is unstable. Holding constant the beliefs-utility benefit at $\lambda B(\rho^*)$, a perturbation that results in, say, a drop in $C(\rho^*)$, would provide $\rho^*$ and some lower neighbors (to the right) the incentive to switch to the fair outcome. 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 $\rho^*$ results in incentives that restore equilibrium. For this reason I restrict attention to stable equilibria.

**Appendix C. The dictator game used to identify subsamples**

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

<table>
<thead>
<tr>
<th></th>
<th>t &gt; 0 (%)</th>
<th>Overall</th>
<th>Endowment</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>High anonymity (N = 69)</td>
<td>48</td>
<td>0.67</td>
<td>0.57</td>
<td>0.84</td>
</tr>
<tr>
<td>Low anonymity (N = 132)</td>
<td>66</td>
<td>1.04</td>
<td>0.85</td>
<td>1.20</td>
</tr>
</tbody>
</table>

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. 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 0.37 higher than in the High Anonymity condition (0.67).

The difference in giving frequency (z = 2.48, p < 0.01) and in mean transfer (z = 2.37, p < 0.01) are both statistically significant. The difference in mean transfers (0.37) is larger than the difference in the mean transfer of dictators endowed with 6 (1.09) versus 6 (0.78), 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.

Appendix D. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jebo.2015.05.008.

References


---

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.