Beliefs-based altruism as an alternative explanation for social signaling behaviors

Vera L. te Velde[*]

University of Queensland

October 10, 2017

Abstract

Social preferences researchers have firmly established the importance of social image to prosocial behavior. Yet an alternative explanation for many such findings has been almost wholly ignored. Instead of, or in addition to, being motivated by maintaining others' good opinion of me, I may be motivated to maintain others' good feelings about how they have been treated. For example, I may give more to a stranger when we meet face to face because they will feel worse being denied in that context than if I denied them anonymously. In an experiment, I find that a substantial fraction of the population does believe that others have beliefs-based utility, and that these second-order beliefs are additionally strongly linked to actual choices in an experimental setting designed to identify BBA by putting monetary utility in opposition to beliefs-based utility. This link establishes the importance of BBA for a large fraction of the population whose actions may be misattributed to social signaling.

JEL classification: C91; D03; D63; D83.

Keywords: Social Image, Social Signaling, Altruism

1 Introduction

"It isn't really about how much money is raised, it is knowing that [those] who are really struggling at the moment, feel that others understand and care."

I recently received an email thanking people for helping out with a small bakesale that ended with this statement. It succinctly expresses a natural motivation for altruistic behavior

[*] vtevelde@gmail.com. I thank Shachar Kariv, Matthew Rabin, Ulrike Malmendier, Julio Rotemberg, Francesca Gino, Lea Heursen, and several seminar audiences for helpful comments. Ishita Arora provided excellent research assistance. All errors are my own. Financial support from the National Science Foundation and the U.C. Berkeley Institute of Business and Economic Research is gratefully acknowledged.

that has not been well studied by economists, even as we've learned a great deal about charitable giving and social preferences over the last twenty years. This is *beliefs-based altruism*. BBA is the desire to make other people *feel* good, independently of the objective, material influence of your actions. A tiny bakesale can't possibly have a noticeable impact on the suffering from a massive drought, but "it's the thought that counts", and the feeling of emotional support that victims get from this type of activism is enough rally individuals around a daunting cause.

The existing literature on prosocial behavior does not attribute this behavior to BBA, however. Instead, in situations where pure and impure altruism can't explain all of prosocial behavior, social image is credited with the remainder. The strength of social image motivations has indeed been endorsed by a mountain of literature. In both the lab and field, we've seen that behavior changes when the anonymity of choices changes (Alpizar, Carlsson and Johansson-Stenman, 2008; Franzen and Pointner, 2012; Soetevent, 2005; Hoffman et al., 1994; Bohnet and Frey, 1999; Andreoni and Petrie, 2004; Andreoni and Bernheim, 2009; Rege and Telle, 2004; Cason and Khan, 1999) and that people prefer to avoid situations where their prosocial (or not) actions will be scrutinized (Dana, Cain and Dawes, 2006; Broberg, Ellingsen and Johannesson, 2007; Lazear, Malmendier and Weber, 2012; Malmendier, te Velde and Weber, 2014). The power of social image has been successfully exploited to increase voter participation (Gerber, Green and Larimer, 2008; Dellavigna et al., 2017), increase charitable giving (DellaVigna, List and Malmendier, 2012), and promote safe sex (Ashraf, Bandiera and Jack, 2014), and been shown to promote teamwork (Babcock et al., 2015; Bandiera, Barankay and Rasul, 2005) and individual work ethic (Mas and Moretti, 2009).

In some of these settings, like Mas and Moretti (2009), the observers are completely passive and unaffected by the decisionmakers who seem to behave differently under scrutiny. Social image thus seems to be a clearcut driving force. In other situations, however, social image may be getting credit for more than it's responsible for, because BBA is tightly confounded with it. The aim in this paper is to take a classic experimental demonstration of social signaling (Andreoni and Bernheim, 2009) and investigate to what extent BBA may lead to similar behavior.

Broadly defined, beliefs-based altruism is altruism over the non-material, beliefs- or information-based aspects of another person's total utility. White lies might be the most familiar example: we tell small lies to friends to preserve their self image, even if a friend appreciates brutal honesty (ruling out a social image explanation), and even though we may realize that they may benefit from the truth (ruling out a simple altruism explanation). In this paper, the beliefs-based preference people are hypothesized to have, and which others are hypothesized to be altruistic towards, is for others to have good intentions toward them. For example, if my friend and I reach for a $10 bill on the sidewalk but he picks it up first and gives me $5, I will have material utility from the $5 but also non-material utility from feeling well-treated. If he had given me $5 from a $10,000 bill, however, I would have had the same material utility, but less total utility due to feeling unfairly treated.

My friend, in this scenario, may of course have many motivations to share with me, even if we were complete strangers who will never see each other again (that is, outside of any repeated game, reciprocity, and/or reputation effects). He may inherently want to do the fair thing. He may think of himself as a fair person and want to preserve his self-image. He may not want me to think less of him. Or, if he has BBA, he may want me to feel like I have been well treated. All of these motivations push in the same direction, and all except the last have been independently studied and distinguished (see DellaVigna, List and Malmendier (2012) and Grossman (2015) among many others.)

I study a modified dictator game in which beliefs-based preferences, and correspondingly beliefs-based altruism, may arise. The modification that allows for different motivations to be distinguished is that a computer may override the dictator's decision with one of two commonly-known allocations, secretly but with commonly known probability. One override is quite biased towards the dictator and the other is (symmetrically, to preserve ex ante fairness) quite biased towards the recipient. Andreoni and Bernheim (2009) found that the introduction of this computer override attracted many dictators to the selfish but veiled option that maintained plausible deniability of selfish intentions. The explanation is that social signaling is a strong motivation, because the chance to be mostly selfish without sending a signal of selfishness is taken by both those who would otherwise be quite generous and those who would share nothing at all.

But BBA could also contribute to this finding. If generous sharers only share generously to avoid hurting others' feelings, then they might take the opportunity to be more selfish when they can do so while still preserving others' beliefs-based utility. And if it's cheap to create beliefs-based utility for others, non-sharers might also share a token amount in order to do so.

BBA and signaling motivations are difficult to disentangle definitively, as discussed further in section 2. But by combining the experiment with a questionnaire on beliefs, I can simultaneously elicit beliefs about beliefs-based preferences and see how these beliefs impact real behavior. I find that while social image is most certainly an important motivation for choosing the veiled selfish option, a substantial fraction of veiled choices may instead be driven by BBA.

The questionnaire elicits beliefs about the magnitude of others' beliefs-based preferences. For example, if someone reports believing that the recipient in this game would prefer to receive the veiled amount than an additional \$1, but would rather receive an extra \$2 than the veiled amount. This would indicate a belief that the veiling of negative information is worth somewhere between \$1 and \$2.[1]. I can then categorize individuals into different types of BBA, relating to the value of good information, bad information, or any kind of information. Establishing that people recognize beliefs-based preferences that others have, however, could be inconsequential if they do not actually modify their behavior accordingly. The crucial next step is therefore to demonstrate a link between types of BBA elicited in the questionnaire and actual choices in the game.

The results show that a substantial fraction (approximately one third to two fifths) of the population believes that others have this form of beliefs-based preference.[2] I also find a strong link between beliefs and behavior, suggesting a somewhat less cynical interpretation of behavior that might appear to be driven purely by social pressure. In particular, I find that a 2nd-order belief that bad information is worth avoiding increases the likelihood of choosing the veiled option in the game, and that a 2nd-order belief that good information is worth seeking increases the probability of generous sharing in the game.

This paper prioritizes careful identification to demonstrate the existence of BBA as definitively as possible. From a scientific perspective, this enriches our understanding of human decision making. More than that, however, this is a necessary first step towards understanding the full importance of BBA in realworld settings. When soliciting charitable donations, for example, an intuitive understanding of BBA can guide the type of emotional appeals presented to potential donors even if the response cannot be ex post definitively attributed to BBA over social image concerns. The results in this paper show that BBA is a plausible enough phenomenon to make these avenues worth pursuing.

[1] This ignores issues relating to interpersonal comparisons of utility, which is not critical to the analysis.

[2] Note that I do not show that people actually have this form of beliefs-based preference. 2nd order beliefs are what are relevant for *altruism* over those possible preferences, which is the potential confound with social image that I am investigating.

2 Identification of BBA

As alluded to in the introduction, BBA and social image are difficult to disentangle experimentally. To understand why this is the case we must define some terminology precisely.

"Image" is a broadly-encompassing term referring to any motivation linked to opinions or inferences about someone. Most straightforwardly, I might be good because I want people to like me. But I also I might be good even if I'm acting anonymously because I want to know that people like me, even if they don't know who I am. Or, I might be good because I want to like *myself*. I might be nicer to people who are affected by my choices because they have stronger opinions of my behavior than disinterested third parties. I might be good to build my reputation, which will help me down the line, or just because I like to be liked.

The terminology distinguishing these concepts is far from standardized, but for the purposes of this paper I will lump them all together under the heading of social image. The key role of social image is that when the emphasis on it increases (due to less anonymity, more transparency, a bigger audience, etc.) the decision maker behaves better, whatever "better" is according to whoever's opinion is motivating the decision maker.

Beliefs-based altruism also potentially encompasses a plethora of subtly different concepts (see section 6) but the flavor of BBA dealt with in this study (because it is most easily confounded with social image) is altruism over utility from intentions. In models such as in Rabin (1993) or Levine (1998), people derive some utility from knowing that others have kind intentions towards them even if this doesn't translate to material utility. For example, I appreciate well-intentioned advice on how to treat a carpet stain even if the cleaner ends up damaging my carpet. Naturally, if I experience positive utility from others' good intentions, others may altruistically act kindly in order to produce those warm feelings.[3]

With BBA and social signaling thus defined, we can see the difficulty in trying to conclusively disentangle them experimentally. This is because not only do they push in the same direction almost always (as do social image and pure outcome-based altruism), they may reinforce each other.

To see this, consider a simple two-person interaction. If Alice shares with Bob, Bob will both infer that Alice is generous and will have a positive emotional experience from

[3] This concept is distinct from, but most closely related to, "guilt aversion" (Battigalli and Dufwenberg, 2007). Guilt aversion is a form of altruism over *reference-dependent* utility, from consumption-based outcomes, determined by a potentially-altruistic decisionmaker. The psychological motivation is also different: avoiding guilt is closer to an image motivation (either self or social) than a desire to make others feel good, which would motivate altruistic acts even in the complete absense of expectations and even if not attributable to the decisionmaker.

his interaction with her. Alice might therefore be motivated to share by either social image or BBA, so they cannot be disentangled. (She might also be motivated by pure or impure altruism, but we already know how to disentangle altruism from image motivations, so I won't reiterate discussion of this issue here.)

What if another observer, Carol, is present? If Alice then shares even more with Bob, can we attribute the change to social image? Not definitively: While it's possible that having an audience for her good deed might reduce the beliefs-based utility Bob experiences because he infers that Alice is only being nice to look good for Carol, it could also be true that failing to be generous in front of Carol would be even worse. Bob would then miserably realize that Alice didn't want to be nice to him *even though* she was being watched by Carol. Alice might reasonably want to spare Bob this misery, leading her to be more generous in front of Carol. What appears to be a response to an audience is actually a response to Bob's inferences involving the audience.

What if Bob is kept entirely in the dark? If Alice is still more generous towards Bob when Carol is watching, can we attribute this conclusively and solely to social image? No: It can be unpleasant or angering to watch others violate social norms, even when we're not involved, and Alice may wish to spare Carol this frustration. Alice's BBA towards *Carol* is easily mistaken for social signaling in this case. We could go even further and try to make social image directly opposed to BBA, but the same problem arises. Suppose, for example, that Bob is a disabled child. Image concerns might motivate kindness towards him, while BBA motivates treating him indifferently in order to save him embarrassment. But if the BBA motive is recognized, acting indifferently could correspondingly lead to a positive image. Similarly, even if we could directly manipulate Bob's beliefs (as we sometimes do when, for example, we reveal anonymous good deeds so that the recipient can experience warm feelings about an interaction with someone who didn't want to reveal or emphasize their own kindness) that revelation itself can be seen as kind and lead to a positive image.

These examples capture the fundamental difficulty: If social image increases in motivational strength, then BBA does too because failing to be nicer when there is a stronger incentive to be nice may be perceived as mean. Conversely, if BBA increases in motivational strength, then so does social image, because responding to the emotions of another can be seen as nice. Social image (jointly with BBA) was simple enough to disentangle from pure altruism because they don't interact in this way; proving the existence of BBA separately from social image is more difficult.

Two routes can be taken forward:

First, we can recognize that some alternative explanations are clearly less plausible than others. Many readers likely scoff at the notion that BBA towards observer Carol could possibly explain Alice's behavior towards clueless Bob. This is the approach taken by the only other experimental attempt at idenfiying BBA: In his dissertation, Grossman (2008) reports on three experiments that plausibly distinguish BBA from social image under some assumptions about what kinds of BBA and image motivations are truly realistic. The results are intriguing, but inconclusive.[4]

Second, we can try to elicit motivations and attitudes more directly and correlate them with behavior. This is the approach taken here. This relies on combining typical revealed-preference evidence from an incentive-compatible experiment with survey evidence asking about beliefs directly. The idea is simple: some people believe that beliefs-based preferences are an important component of others' utility, and some don't, and it is straightforward to categorize people accordingly based on survey data. If the former group behaves in ways more likely to communicate good intentions than the latter, then BBA is the likely source of the difference. The details are presented in the following section. This approach reveals a strong link between beliefs in the value of information and choices that are consistent with altruistic concern over those beliefs-based preferences.

3 Experiment design

The experiment consisted of three parts that each participant completed in some order.

- *Game*: Each subject participated in a one-shot variant of a modified dictator game. Each subject makes a decision contingent on being assigned the role of dictator, and roles are randomly assigned ex post. As dictator, each participant first chooses a quantity x_R between \$0 and \$1 to share with the recipient, and keeps the remainder $x_D = 1 - x_R$. This choice is then randomly overridden with commonly-known probability p, unobserved by the recipient. With probability $p/2$ the allocation

[4] Each experiment relies on a third party "informer" to display BBA, following the intuition mentioned above that someone with BBA may wish to incur personal cost to positively affect the beliefs of someone else. The results find little evidence of this, which is certainly intriguing, but interpreting this as evidence against BBA relies on the assumption that the informer doesn't have any social preferences towards the decisionmakers themselves. Even though the decisionmaker is not informed about whether their choice was revealed, the informer may feel that they normatively shouldn't reveal someone else's decisions without their knowledge or consent. It also relies on the assumption that there is not a negative social image (or negative beliefs-based utility!) associated with revealing others' private acts. Conversely, if many informers *had* paid to reveal information, even this can't prove that BBA is the motivation, because once again, there might be a positive social image associated with incurring a cost to make someone else feel better.

$(x_D, x_R) = (.91, .09)$ is chosen, and with probability $p/2$ the allocation $(.09, .91)$ is chosen. The randomly imposed default thus preserves the 50-50 split as the ex ante fair choice.

After making this choice, each subject also makes an incentive-compatible guess of the average choice made by other subjects. This serves both to incentivize paying attention to the details of the instructions and as a measure of expectations. Example instructions for this game are provided in the Appendix.

This game is modeled on the one first implemented by Andreoni and Bernheim (2009), who find, in short, that people will take the option to be selfish while preserving their social image, by choosing the veiled option when the probability of a computer override is high enough that that outcome can't easily be attributed to the decisionmaker. But I show that choices in this game are also related to elicited BBA, as measured in the questionnaire:

- *Questionnaire*: Each subject answered 10 questions about a hypothetical game played between two individuals, Alice (the dictator) and Bob (the recipient). Alice and Bob are described as playing the same game as above, but with a total allocated amount of \$10 and potential computer-set allocations of $(0.87, 9.13)$ and $(9.13, 0.87)$. Subjects first state their opinion of what Alice would do in this situation. They then make an incentive-compatible guess of the average answer to the previous question among all participants. This again serves to incentivize paying attention to the details of the setting along with empathic consideration of others' experience with the setting, which is required for the later questions.

  Each subject then answered 8 questions about their opinions on the relative merit of several possible choices. These form the key measure of BBA. For two values of ϵ per person, each participant is asked whether it is "better" (where "better" is defined below in section 3.1) to choose \$0.87 than \$0.87$\pm\epsilon$ and whether it is better to choose \$9.13 than \$9.13$\pm\epsilon$. For example, with $\epsilon = 5$, a subject might be asked whether Bob would be happier after receiving \$0.82 or \$0.87, and whether he would be happier after receiving \$0.92 or \$0.87, and similarly for \$9.08/\$9.18 versus \$9.13.

  Because opinions like these can't be incentivized to be truthful, data quality is a concern. This is discussed in detail in section 5.1. Several design measures mitigate this concern: The order of possible responses is randomized, and subjects are instructed that they must answer honestly in order to be eligible for bonus payments, by avoiding contradicting themselves.[5] Because it is easy to inadvertently self-contradict if not taking care, and because it is easy to detect contradictory responses, it is possible to estimate rates of sloppy responses. There is no indication that responses aren't reliable (see section 5.1), and issues with interpretation are addressed as they arise.

Example instructions for the questionnaire are included in the Appendix.

- *Survey*: Following the game and questionnaire, each participant answered a brief 5-question survey to provide basic demographic information about age, gender, income, education, and race. The survey is included in the Appendix.

3.1 Treatments

I employ a $2 \times 2 \times 2 \times 3$ treatment design, although the analysis primarily uses the first two dimensions of variation.

- *Game first or second*: The order of the game and questionnaire components was randomly assigned. The demographic survey always followed these two components. This was done in order to measure baseline responses in both the game and questionnaire uninfluenced by previous responses to the other, perhaps due to either familiarity, motivated reasoning, or a desire to be consistent.

- *Small or large values of ϵ*: Each value of ϵ generates four opinion questions on the questionnaire, and each participant is assigned two different values of ϵ, for a total of eight questions. Participants in the small values treatment were assigned $\epsilon = 5¢$ and 40¢, and participants in the large values treatment were assigned $\epsilon = 20¢$ and 70¢. This was done to keep the survey short for each individual while generating the desired variation and also to verify that responses are sensitive to the values used.

- *High or low social pressure*: A smaller value of p means that the dictator's decision is more directly observable by the recipient. Subjects were randomly assigned to either a high social pressure treatment, in which p in both the game and questionnaire was 10%, or a lower social pressure treatment with $p = 50\%$. This treatment dimension is modeled after Andreoni and Bernheim (2009); however, no significant social pressure effects were observed in the highly anonymous online setting.

[5] However, roundabout rationalizations can be generated for almost any combination of answers, so no participant was actually deemed ineligible for payment based on their responses.

- *Question phrasing*: The main questionnaire treatment asked participants questions of the form “Do you think Bob would be happier after receiving a payment of \$0.87 or \$0.92?”. Two other treatments (used as manipulation or placebo checks and discussed in section 5) varied this phrasing to be either “Do you think Bob would prefer for Alice to choose to share \$0.87 or \$0.92?” or “Do you think it’s more morally appropriate for Alice to choose to share \$0.87 or \$0.92?”. I refer to these three phrasing treatments respectively as “ex post happiness”, “ex ante preference”, and “moral appropriateness”.

Except as discussed in section 5, all analysis that follows uses only the “ex post happiness” phrasing treatments.

Additionally, due to a technical problem that was rectified after the experiment began, data for the four questions based on a high-ϵ in the questionnaire was not recorded for some subjects. This did not impact the subject experience, however, and is thus not analyzed as a separate treatment. The impact of this issue is discussed in relevant results.

3.2 Recruitment

The experiment was implemented on a private webserver and recruitment was done through Amazon’s online workforce, Mechanical Turk or MTurk. A link to the experiment was posted as a Human Intelligence Task, or HIT, that any eligible user of the website could accept. Participation was restricted to MTurk account holders in the United States. I additionally discarded data from users with IP addresses outside the U.S., multiple submissions from the same IP address,[6] and partial (and unpaid) submissions. Out of 2468 total submissions and partial submissions, 162 were dropped from analysis for one of these reasons.

The advertizement listed a fixed payment of \$0.40 USD in addition to bonus payments up to \$1.10 contingent on choices and on chance. The advertizement also stated an estimated completion time of less than 5 minutes and imposed a time limit of 15 minutes. Two subjects took slightly more than 15 minutes to complete the experiment but were paid manually and their data retained normally. Subjects followed a link to an external private server, which assigned the user to a treatment group and recorded their responses. After submitting the survey, the subject was provided with a key value to enter into MTurk in order to received payment automatically from Amazon. A screenshot of the MTurk posting is provided in the Appendix.

[6] IP addresses were collected, then used to generate state-level geographic information for each participant, and then permanently deleted prior to matching geographic data to response data in order to preserve anonymity.

Random assignment to treatment groups was based on arrival time; sequential participants were assigned to sequential treatment groups. At the end of the experiment, additional assignments to specific treatment groups were made manually to ensure an even number of participants in each group, which was necessary due to a handful of incomplete submissions. Subjects were also paired in dictator/recipient groups based on arrival time; partners were always sequential participants in the same treatment group. Partners were not required to interact in realtime since no responses to others' actions were required and payments were calculated ex post.

Additional analysis of the recruitment methodology is reported in te Velde (2016).

3.3 Payment

Participants were paid a fixed $0.40 for completing the study. This was the piecerate specified on the HIT which was thus automatically paid by Amazon unless I actively rejected their submission as valid. Incentivized payments were paid as an additional bonus payment, which any HIT poster can make in any amount to any person who has completed a HIT by that poster. This flexibility in MTurk is what allows implementation of classic incentivized economic experiments, as reported by Horton, Rand and Zeckhauser (2011), Mason and Suri (2010), Paolacci, Burson and Rick (2011), and Chandler, Mueller and Paolacci (2014).

In this experiment, as described above, potential bonus payments came from three sources: subjects were paid $0.05 for guessing within 10 cents of the median shared quantity in the dictator game, and were paid $0.05 again for guessing within 25 cents of the median response to the question asking what you think Alice would do in the hypothetical dictator game.

Lastly, subjects were probabilistically paid for the outcome of the dictator game. The dictator and recipient roles were randomly assigned after both made their dictator choices, and the result of that game was reported to both subjects in a message that accompanied their bonus payments.[7] Then with a 20% probability, they were paid according to the result.

Overall, 2306 subjects were paid an average of 53 cents (median $0.45) for an average of 4 minutes and 38 seconds (median 3:59) of participation. These stakes, while of course smaller than similar (but much longer) laboratory studies, are substantially higher than the median wage rate on MTurk. Various authors have found little impact of small stakes on outcomes in social preferences experiments on MTurk (Amir, Rand and Gal, 2012), and Bohannon (2011)

[7] Incidentally, messages cannot be sent in the absence of a bonus payment, so every participant was paid a minimum 1¢ bonus. They were not informed of this at the time of the experiment.

in fact suggests that higher stakes can decrease the quality of data obtained on MTurk, so it is highly doubtful that the low stakes are crtical for the results; this is discussed further in section 5.1.

4 Results and Discussion

Sections 4.1 and 4.2 report on demographics and randomization and the replication of the (MTurk adapted version) of Andreoni and Bernheim (2009). Section 4.3 discusses the main novel results of the paper that result from combining the questionnaire data with the game data, and section 5.2 examines the possibility that these results may be due to motivated reasoning rather than BBA.

4.1 Demographic statistics

Table 1 shows demographic summary statistics taken from the survey data and by tracking IP location and activity on the experiment website. The demographic statistics are broadly in line with previous findings on the U.S. MTurk subject pool (Ipeirotis, 2010) and are overall well-randomized across treatments; details are omitted but can be replicated with the data provided online. These demographics are used as controls in the analysis to follow; gender or racial group occasionally attains statistical significance but no demographic variable influences the coefficient estimates of interest.

4.2 Game results

The game phase asks two questions: how much do you choose to share with your partner in this game (given that you are chosen to be the decisionmaker and that the computer doesn't override your choice), and how much do you think your partner will choose to share with you? Figure 1 shows sharing behavior by treatment in the game phase; approximately half of participants chose to share exactly half of the $1 endowment, and among the rest, social pressure levels had little effect on behavior while the order in which the game and questionnaire had substantial effects, as expected. The questionnaire makes salient the BBA-based reasoning for choosing the veiled option, and correspondingly participants who play the game after answering the questionnaire are more than twice as likely to do that. This increase seems to come at the expense of more generous sharing.

Table 1: **Participant Demographics**

	Mean	**Lower Bound**	**Upper Bound**
Female?	0.421	0.400	0.441
Under 30?	0.628	0.608	0.648
30s?	0.223	0.207	0.241
40s?	0.075	0.065	0.087
50s?	0.053	0.045	0.063
60s+?	0.020	0.014	0.026
High school or less?	0.104	0.092	0.117
Some college?	0.454	0.434	0.475
Bachelors?	0.339	0.320	0.359
Post graduate?	0.094	0.082	0.107
Household Income <$20k	0.233	0.215	0.251
Household Income <$50k	0.359	0.339	0.379
Household Income <$80k	0.232	0.215	0.250
Household Income <$150k	0.141	0.127	0.156
Household Income >$150k	0.035	0.028	0.044
PDT timezone?	0.189	0.173	0.206
MDT timezone?	0.064	0.055	0.075
CDT timezone?	0.242	0.225	0.261
EDT timezone?	0.499	0.479	0.520
White?	0.783	0.765	0.799
Hispanic?	0.032	0.025	0.040
Black?	0.059	0.050	0.069
Asian?	0.092	0.081	0.105
Seconds spent on survey	15.872	15.311	16.433
Seconds spent on game	99.619	87.487	111.750
Seconds spent on questionnaire	162.698	158.488	166.908

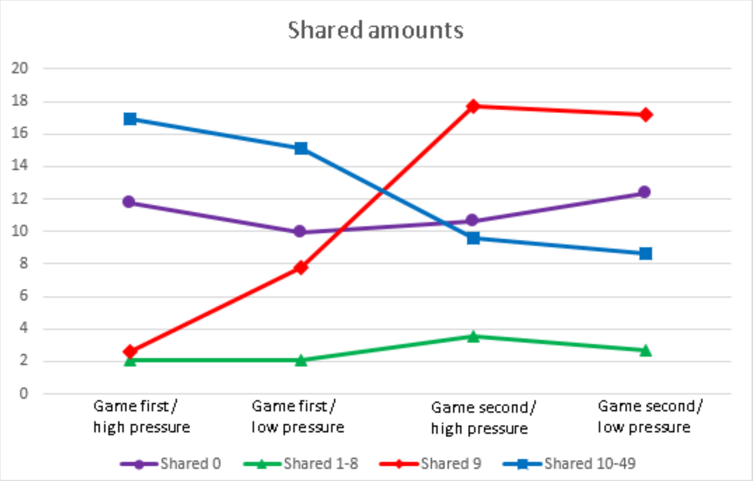
Bounds are 95% confidence bounds for binomial or normal variables.

Table 2 reports regression results. Controls used are dummies for gender and race and indices of age, income, and education corresponding to the survey responses. Results without demographic controls are omitted for brevity but are extremely similar both in magnitude and significance.

Column 1 shows that rates of sharing are high, at an average of 42¢ per dollar in the reference treatment (low pressure, game first). Columns 2 and 4 shows that beliefs are strongly related to behavior; those who guess that others are more generous also behave much more generously. Column 3 shows that the average guess is less than 1 cent from the true average shared amount.

More interesting than aggregate sharing levels is the choice to try to veil a relatively selfish decision by going with an option that can be plausibly blamed on a computer override.

Figure 1: Game results summary



Frequency of sharing choices by treatment in the game phase of the experiment. Choosing to share 9 cents exactly corresponds to the more selfish of two veiled options that can be plausibly blamed on a computer override. Approximately half of subjects chose to chare exactly 50% and are omitted from the figure.

Constants aren't reported in columns 5 through 8 since the table shows marginal effects, but in the reference treatment 7.8% of participants choose to share exactly 9 cents, the "veiled" option that can be plausibly blamed on a computer override. Expectations were again accurate on aggregate: 8.3% of participants guessed that their partner would share 9 cents.

Columns 6 and 8 show that as in overall sharing levels, veiled behavior is also tightly correlated with beliefs about others' veiled behavior.

The questionnaire naturally makes the veiling option much more salient and can even be said to frame it as a kind option by indirectly highlighting the potential beliefs based altruism motivation for doing so. The second and third rows of table 2 examine this order effect. As expected, it causes people to be much more likely to choose the veiled option and correspondingly causes the overall level of sharing to drop. Beliefs follow suit but even moreso: people believe that the salient veiling option will be taken advantage of much more often than it actually is.

Social pressure does not have a significant impact on either overall sharing behavior or expectations of sharing in this game (columns 1 through 4), which is not too surprising given that even the high social pressure treatment is played by anonymous strangers online. I am therefore, as expected, not able to directly compare these results to those of Andreoni and Bernheim (2009) (who use a particularly strong form of social pressure), but I can corroborate the suspicion that social pressure is a weak force in online settings. This indirect evidence of the weak power of social pressure in this online setting further supports the notion that veiled choices are more likely to be driven by BBA in this setting.

Table 2: **Game results**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	Shared	Shared	Guess	Guess	Shared 9¢?	Shared 9¢?	Guessed 9¢?	Guessed 9¢?
High Social Pressure	1.324	3.377	-4.648	-5.352**	-0.0739**	-0.0544*	-0.0622	-0.0300
	(2.731)	(2.480)	(2.955)	(2.685)	(0.0374)	(0.0313)	(0.0461)	(0.0454)
Game After Questionnaire	-6.502**	-2.091	-9.985***	-6.529**	0.0926***	0.0137	0.191***	0.153***
	(2.743)	(2.538)	(3.113)	(2.824)	(0.0298)	(0.0247)	(0.0372)	(0.0370)
High Pressure × Game After	2.372	-0.278	5.999	4.738	0.0735	0.0594	0.0511	0.0188
	(3.863)	(3.381)	(4.243)	(3.690)	(0.0563)	(0.0497)	(0.0616)	(0.0566)
Guess		0.442***						
		(0.0353)						
Shared				0.532***				
				(0.0427)				
Guessed 9¢?						0.389***		
						(0.0522)		
Shared 9¢?								0.544***
								(0.0605)
Constant	42.62***	23.23***	43.89***	21.24***				
	(1.994)	(2.464)	(2.147)	(2.541)				
Observations	736	736	736	736	736	736	736	736
Controls	Y	Y	Y	Y	Y	Y	Y	Y

Standard errors are robust and reported in parentheses. Coefficients significantly nonzero at .99 (***), .95 (**) and .90 (*) confidence levels. All controls are demeaned such that constant represents overall averages within the reference (low social pressure, game before questionnaire) treatment. Columns 5-8 are estimated using probit regression and marginal effects reported, or the effect of a change in an indicator variable from 0 to 1.

4.3 Questionnaire results

The questionnaire phase of the experiment asks participants to answer 10 questions. First they are asked what they think a hypothetical person would do in the game, and then what they think other participants think a hypothetical person would do in the game. Responses larger mirror those of the game phase described above. Social pressure does not have a substantial effect on behavior or guesses, and experience from the game, if played prior to taking the questionnaire, doesn't substantially impact guesses in the questionnaire.

The remainder of the questionnaire, however, is of primary interest. These are the eight questions that ask participants about their beliefs about the ex post happiness of others in various circumstances. Recall that \$0.87 and \$9.13 are the veiled options, the former of which can be chosen to achieve a selfish allocation while maintaining plausible deniability. The 8 questions simply ask whether \$0.87 or \$9.13 is a preferable *outcome* from the perspective of the recipient than \$0.87$\pm\epsilon$ or \$9.13$\pm\epsilon$.

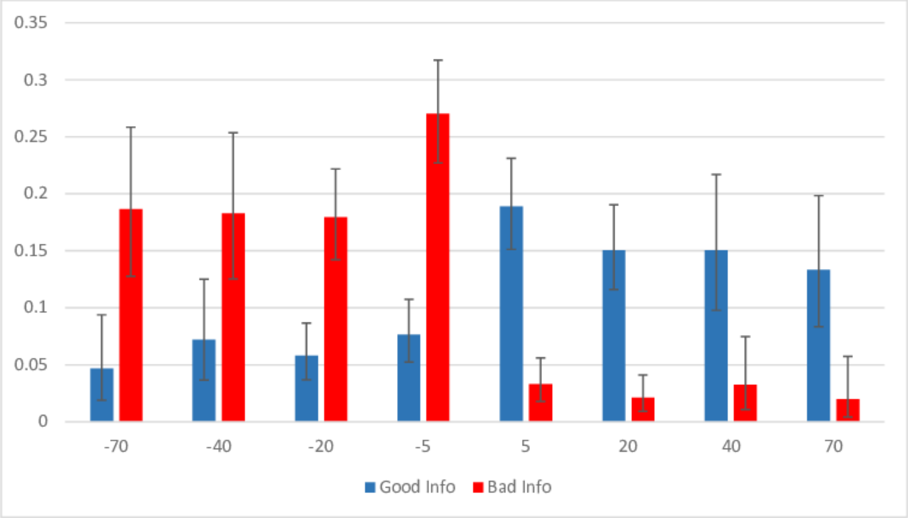
These questions measure 2nd order beliefs of beliefs-based preferences. Believing that someone prefers \$0.82 to \$0.87 is the belief that someone prefers to sacrifice 5 cents in order to gain certain information about the source of the money. This information is "bad" in that it confirms that someone has treated you selfishly, and so this corresponds to a belief that bad information is and worth at least 5 ¢. Similarly, believing someone prefers \$0.87 to \$0.92 indicates a belief that bad information is worth at least five cents to *avoid*. Similarly, stating that \$9.08 is preferable to \$9.13 indicates that good information is worth at least five cents, and stating that \$9.13 is preferable to \$9.18 indicates that good information is worth at least five cents to avoid. Note that BBA can only be indicated by a statement that a smaller amount of money is preferable to a larger amount in some instance.[8]

Figure 2 summarizes responses across treatments. The four bars at ±5 correspond to the fraction of subjects who answered the example questions described in the previous paragraph in a way that indicates a strictly positive or strictly negative value of information. The bars at other values measure the frequency of similar responses to questions comparing \$0.87 or \$9.13 to \$0.87$\pm\epsilon$ or \$9.13$\pm\epsilon$, with $\epsilon = 20$, 40, or 70¢. It's clear that a substantial fraction of respondents believe that others prefer to trade money for a lack of bad information. Fewer, but a still substantial fraction, conversely believe that others prefer to trade money for good

[8] From this point on, rather than pedantically referring to 2nd order beliefs about beliefs-based preferences, I will refer simply to BBA, on the assumption that if someone believes that someone else prefers A to B, they will give them A rather than B if the price is small enough. Under the additional assumption that people weight their own utility higher than others, these measures of 2nd order beliefs about beliefs-based utility thus represent upper bounds on the extent of BBA.

information. A handful of people also state that (they believe that others feel that) bad information is worth paying for, and a negligible fraction state that good information is better avoided.

Figure 2: Value of Information



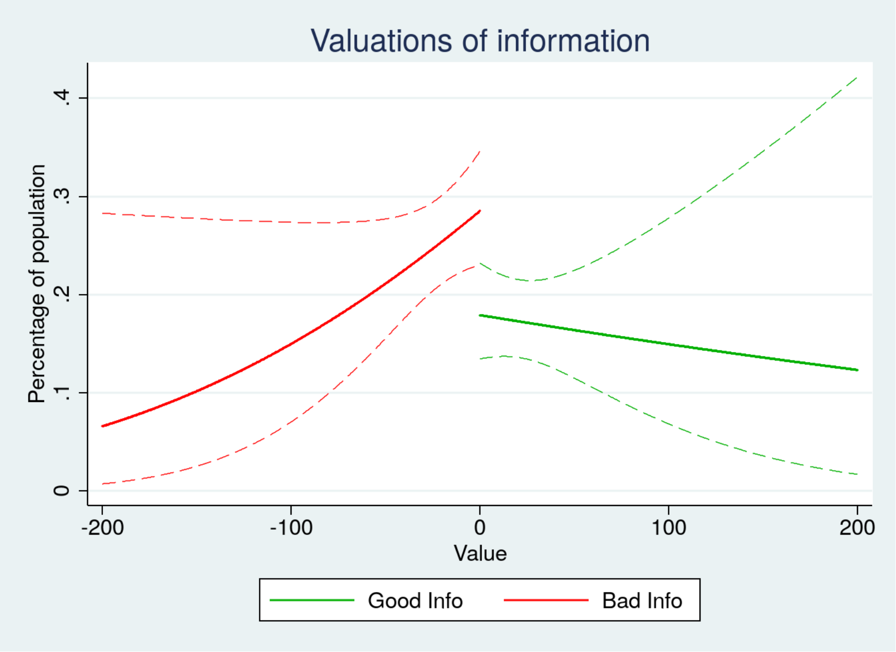
The percentage of respondents who believe that either good or bad information about someone's treatment of them is worth at least the given amount, in cents. Negative values indicate a belief that that type of information is worth something to be avoided.

Figure 3 extrapolates from these data to show the estimated fraction of the population that believes that others value four types of information at each minimum absolute value. This function is based on a probit estimation of the likelihood of stating the given value of information as a function of the monetary trade-off and uses only questionnaire responses from subjects who had not yet played the game (although combining all treatments yields similar estimates).

Another breakdown of this data categorizes people by the type of 2nd order beliefs about beliefs-based preferences they state. Figure 4 shows the four most prominent groups in this categorization.

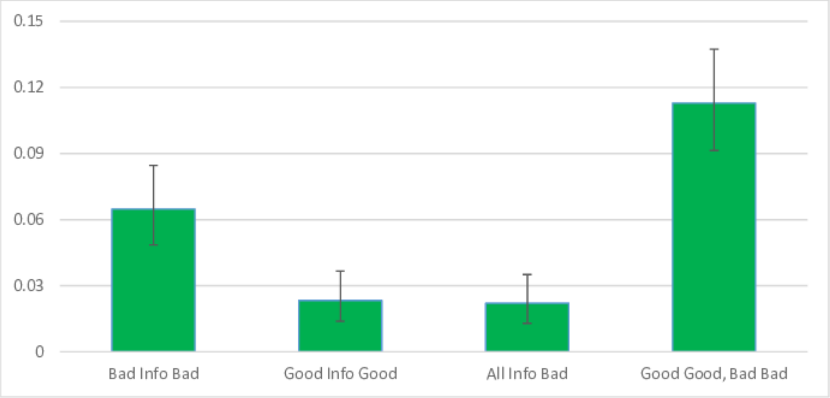
By far the most important category is, as expected from figure 2, those who think bad information is something to avoid inflicting on others and that good information is something worth pointing out to others. Some additionally believe the former but not the latter, and a few believe the latter but not the former. A few more additionally believe that all information is worth avoiding, perhaps in part because explicitly choosing to reveal very generous behavior could be interpreted as bragging (see also section 5.1). Overall a little

Figure 3: Value of Information



The estimated percentage of respondents who believe that good information is worth at least a given amount, or that negative information is worth at least the given absolute value to avoid. Estimates are based on a probit regression of stated BBA on the relevant value, and only "game second" treatments are used to avoid influencing the estimates with experience in the game. Including "game first" treatments results in extremely similar estimates but with narrower confidence intervals.

Figure 4: Types of Beliefs Based Altruism



The percentage of respondents who express BBA of various types, based on a cost of information of either 5¢ or 20¢.

less than a third of the population expresses some form of beliefs based preferences that exceeds the minimum monetary value inquired about in their treatment - this is of course an underestimate of the fraction of the population who perceives beliefs-based preferences of any degree.

4.4 Beliefs are linked to choices

Subsection 4.3 establishes that people recognize that others have beliefs-based preferences, but that on its own merely implies that someone with any degree of altruism would also exhibit BBA. But by combining these survey responses with the choice data we can see that choices are in fact strongly linked to stated beliefs about beliefs-based preferences, clearly demonstrating the relevance of BBA to choices in this experimental setting.

Table 3: Stated BBA and choices are linked

	(1)	(2)	(3)	(4)
Dependent Variable:	Shared 9¢?		Shared 50¢?	
Bad Info Bad	0.149***	0.191***	-0.0304	-0.0868
	(0.0426)	(0.0617)	(0.0543)	(0.0756)
Good Info Good	-0.0398	-0.0671**	0.112*	0.157*
	(0.0258)	(0.0285)	(0.0582)	(0.0826)
High Social Pressure		-0.0203		-0.0236
		(0.0267)		(0.0432)
Bad Info Bad × High Pressure		-0.0495		0.114
		(0.0385)		(0.105)
Good Info Good × High Pressure		0.0991		-0.0924
		(0.0974)		(0.118)
Observations	736	736	736	736
Controls	Y	Y	Y	Y

Standard errors are robust and reported in parentheses. Coefficients significantly nonzero at .99 (***), .95 (**) and .90 (*) confidence levels. All three columns are estimated using probit regression and marginal effects reported, or the effect of a change in an indicator variable from 0 to 1.

The numbers of people who express BBA of the forms indicating that bad information is valuable or that good information is harmful are small and so the regressions in this table don't distinguish these as separate groups. But the large numbers of people who believe bad information is bad, or that good information is good, or both, clearly behave differently in the game than the rest of the population. Say that a person is a BIB (Bad Info Bad) type if they indicate that bad information is worth a small amount of money to avoid, and that a person is a GIG (Good Info Good) type if they indicate that good information is worth a small amount of money to hear. These are the types used as explanatory variables in table 3.[9]

Columns 1 and 2 show that believing bad information is harmful strongly increases the likelihood of choosing the veiled option. This is a large magnitude effect, strongly significant, and robust to various other specifications. Intriguingly, the other common type of BBA, the belief that good information is valuable, also seems to have an impact on choices, somewhat decreasing the likelihood of choosing the veiled option. Columns 3 and 4 investigate whether this is driven by the desire of these types to choose an even more generous option and therefore send good information to their partner. The effect isn't quite as strong as the bad information channel, but indeed believing in the value of good information does seem to be associated with adhering to a 50-50 split norm.

Columns 2 and 4 investigate whether social pressure interacts with these mechanisms through a change in the goodness or badness of information that outcomes represent. But just as social pressure had little impact on behavior or beliefs overall, it also has little impact on those who are particularly attuned to beliefs-based preferences.

5 Discussion

The interpretation of these results hinges on a few issues: The quality of unincentivized survey data, the possible confounding channel of motivated reasoning, and possible misinterpretations or alternative interpretations of the survey questions. Here I discuss why each of these factors is unlikely to be driving the results.

5.1 Survey data

Since I do not elicit beliefs about objectively verifiable values, the questionnaire half of this experiment is necessarily unincentivized. However, the questionnaire design mitigates these concerns a priori and also allows me to detect careless responses, which are quite low and cannot explain the results.

[9] These types labels ignore some information. First, it combines the "small values" and "large values" treatments so that individuals who are BIB or GIG types, but only at smaller monetary values than they were asked about, are not treated as such. The results are extremely similar if the two treatments are analyzed separately, however, so they are pooled for the sake of brevity. Second, it ignores the responses to the question at a higher monetary value. Categorizing people according to combinations of responses that are not equally likely under random choice introduces differential bias and noise, so only single questions are used as explanatory variables in the regression analysis.

There are in fact a number of reasons to suspect ex ante that the data would be reliable: First, the questions are about preferences of other people in a contrived environment, so it's hard to think of a source of sensitivity or a suggested "right" answer that would cause people to prefer one response over another systematically. Second, the order of options was randomized so anyone answering carelessly should only increase the noise of the dataset, leading to an underestimated effects. Third, the questionnaire stated that participants were only eligible for bonus payments if their responses were not self-contradictory, and that this could be easily ensured by answering the questions carefully and honestly.[10] Fourth, the unincentivized questions come after an incentivized question that is designed to make respondents seriously contemplate the scenario they are being asked about.

But more important than these a priori reasons is the evidence of data quality that comes from the data itself. First of all, the fourth treatment arm that is briefly discussed in section 5 changes the phrasing of the questions and leads to a large change in responses in the intuitive direction, indicating that the phrasing used is conveying the intended meaning and that respondents are answering accordingly. Secondly, the second treatment arm varies the monetary values that subjects are asked to compare to, and as can be seen in figure 2, subjects do in fact change their responses in the intuitive direction based on the values they see. Thirdly, and most critically, incoherent responses are easy to detect and very rare in the dataset. This is because two types of inconsistencies are possible: someone might, for example, state that good information is worth 70 cents but that it's not worth 40 cents. Or, they might state that it's worth 40 cents and also that it's worth 40 cents to *avoid.* Participants answered a set of 8 questions, and among the 2^8 possible answers, a full 75% contain one of these inconsistencies. Even among the participants for whom four responses were lost due to the technical glitch mentioned in section 3, half of the possible responses contain one of these inconsistencies. But in the data, a mere 5% of subjects express incoherent preferences.

This indicates that any careless responders are, at worst, contributing a small degree of noise to the data. The one bias that may be introduced is in the categorization of different types of BBA, shown in figure 4. Random responses are not equally likely to be categorized as each of these types; however, the likelihood of a random response being categorized as "Good info good and bad info bad" is $\frac{1}{64}$ and the likelihood of being categorized as "Good info good" or "Bad info bad" is $\frac{1}{128}$, so this difference is dwarfed by the actual

[10] However, since creative rationalization could pseudo-plausibly justify nearly any set of answers, I ultimately chose to keep every participant eligible for bonus payments.

magnitudes. Additionally, since at most around 8% of subjects are answering carelessly (based on 5% incoherent responses and a 60% likelihood of carelessness being detected in a random response), miscategorization of these subjects would be barely perceptible.

In the game itself, the small stakes sizes could also be seen as a cause for concern. While stakes sizes haven't been shown to be critical to MTurk replications of social preferences experiments, as mentioned in the experiment design, MTurk has only recently grown in popularity within economics and skeptics remain many. Despite the fact that MTurkers skew towards low income and the payment rate in this experiment was substantially higher than what they can usually earn with other HITs, it certainly seems likely that the unusually high rates of sharing fully half of the dictator game endowment can be attributed to the small stakes. But notice that this merely reduces the motivations to engage in social signaling, or to try to be selfish with plausible deniability, relative to laboratory implementations of the same experiment. This is also the difficulty with comparing these results directly to the original Andreoni and Bernheim (2009) implementation, which focuses on the effect of extreme social pressure; fortunately, this comparison is irrelevant to establishing a link between elicited beliefs and behavior, which is the goal of this experiment. Details of the setting including the stakes and the social environment would indeed change behavior and change the fraction of the effect that is due to BBA relative to social image, but would not change the fact that BBA is in fact responsible for some fraction of the effect.

5.2 Motivated reasoning

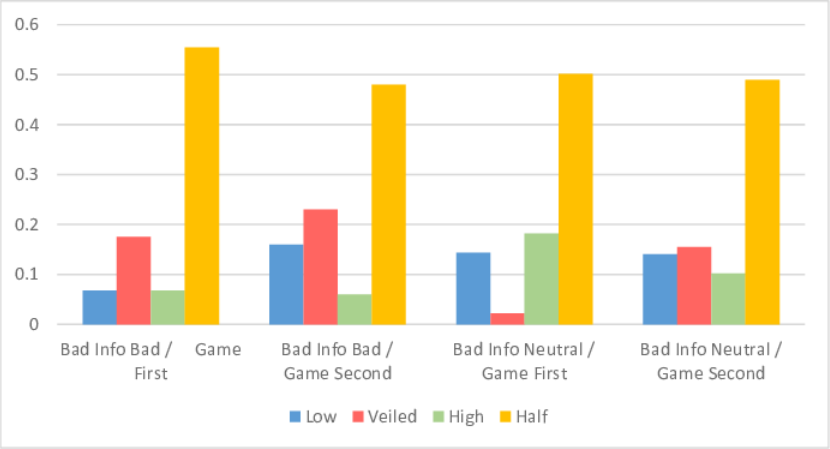
The relationship between GIG BBA and exactly equal allocations cannot possibly be the result of motivated reasoning, but it's possible that the pattern of BIB BBA promoting the veiled choice could be explained that way. Perhaps people are not truly motivated to spare the feelings of others but are using this logic as a chance to be selfish while still telling themselves that their actions are considerate. Evidence of such convenient rationalization for the purpose of avoiding altruism has been shown by, e.g., Di Tella et al. (2015) and Andreoni and Sanchez (2014), so it would not be surprising if similar rationalization occurred here.

The perfect test to address this question is unfortunately not possible: We want to know how people who choose the veiled option and then indicate that they are BIB types *would have* been categorized if they had answered the questionnaire prior to the game. Some tentative evidence is available, however. Figure 5 shows how BIB and non-BIB people play the game before or after the questionnaire, taking this BIB categorization at face value and imagining for the moment it isn't affected by prior game play. There are a couple noteworthy trends in this figure.

First, it appears that those who don't think bad information is harmful are much more likely to choose the veiled option after taking the questionnaire, while BIB types are only slightly more likely to do so, even though these are precisely the people who have given themselves a free excuse to be selfish. This suggests that the questionnaire, more than providing BIB types with an excuse to be selfish, is making salient the opportunity to be selfish with plausible deniability to those who admit to thinking this is not an altruistic thing to do. This is further indicated by the drop in "high" (but not quite equal) sharing choices among those who are not BIB types but play the game after the questionnaire: these are individuals who profit from the veiled option, unlike those who share almost nothing but might be persuaded to share 9¢ if the value of information is pointed out to them.

Second, looking at the proportions of people who choose to share less than 9 versus exactly 9 in the "game first" treatment, we see that the ratios are essentially reversed among BIB types and non-BIB types. This indicates that many of the people who share 9 cents and then report being BIB are doing so honestly, because it would have been cheaper and still self-consistent to share nothing and report that doing so does not inflict additional negative beliefs-based utility on the recipient. There is a drop in high sharers in the BIB/game first group, compared to the non-BIB/game first group, but this is offset by an increase in even more generous equal sharers.

Figure 5: Sharing as a function of BBA type

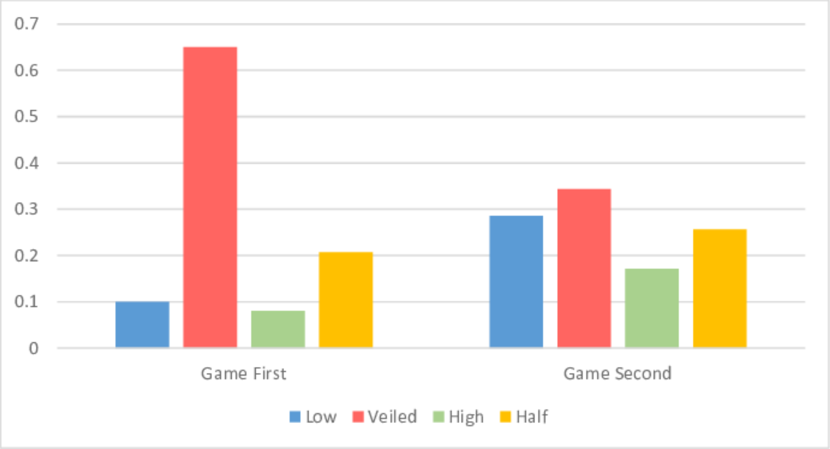


The percentage of respondents who share different quantities, broken down by the "game first" treatment dimension and by expressed BBA type with respect to bad information. Low = 0-8¢, veiled = 9¢, high = 10-49¢, and half=50¢, out of $1 total.

Inversely, figure 6 shows how BIB and GIG categorizations change based on gameplay, taking gameplay at face value and imagining for the moment that it isn't affected by previous categorization.

The glaring trend in this figure is that those who have shared the veiled amount are very likely to then be categorized as BIB, each category of giving in the game, when played after the questionnaire, has similar fractions of BIB types despite the fact that being categorized as BIB in the questionnaire gives these people, and *only* these people, a ready rationalization for choosing the veiled quantity in the game.

Figure 6: BBA type as a function of sharing



The percentage of respondents who are categorized as thinking bad information is worth avoiding, according to their choices in the game and whether they played the game before or after answering the questionnaire. Sharing choices grouped as: Low = 0-8¢, veiled = 9¢, high = 10-49¢, and half=50¢, out of $1 total.

Overall, it seems that the order effects of game vs. questionnaire are more pronounced among the population that is not categorized as BIB and/or shares the veiled quantity. If the order effects are primarily about salience of the veiled option, this makes perfect sense since those who choose the veiled option in the game are clearly already aware of the advantage of this choice, while everyone else may not be and may be more strongly influenced by the questionnaire. This doesn't entirely rule out motivated reasoning, but does suggest that other reasons for veiled sharing, including honest BBA, are more important. And in a way this is expected: "at least I spared their feelings" may be a good way to justify sharing 9 cents instead of 0 or some other very small quantity, but is not a credible way to justify sharing a small quantity in the first place.

Finally, recall that the trend of GIG types sharing more generously is evidence of the importance of BBA that *can't* be explained with motivated reasoning (just like the apparent shift from sharing nothing to sharing the veiled amount among BIB types mentioned above).

So while this relationship is not quite as strong as the link between BIB type and sharing the veiled amount, it is somewhat purer evidence of BBA in that it is not subject to this confound.

5.3 Interpretation

This experiment was a deliberate trade-off in favor of clear identification at the price of being able to directly infer the relative importance of BBA in realworld circumstances. The unfortunate downside to this approach is that the setting used is quite *unlikely* to produce strong effects of BBA, so that the results may be interpreted more pessimistically than they should.

The main reason to believe BBA may be more relevant in natural settings in which BBA and social image and impossible to clearly disentangle is that "good" and "bad" information in this context is ambiguous. First of all, "good" information in this context is potentially colored in its interpretation by bragginess. That is, receiving a very generous share that is revealed to be generous by the giver conveys a different meaning than a veiled generous act that is, for example, revealed by a third party. This is the appropriate comparison for discerning potential beliefs-based altruistic explanations for behavior that might appear to be due to social image motivation, but "good information" is an imperfect description of the information conveyed and the concepts shouldn't be confused. Conversely, "bad" information may be partially viewed as an honest admission on the part of the dictator. These two alternative interpretations of information may be why approximately 8 percent of people think that good information is worth avoiding, and that approximately 3 percent of people think bad information is worth seeking.

This tension between conflicting notions of good and bad also arises when interpreting the survey questions themselves. The fourth treatment arm that varied the phrasing of the survey questions sheds some light on the situation. This treatment arm was intended to ensure that the primary phrasing treatment, analyzed in depth above, captured the intended beliefs. The dramatically different results in the other two phrasing treatments are reassuring to that end. But they are also interesting in their own right.

The primary "ex post happiness" phrasing asked which outcome do you think the receiver would be happier with. The 2nd "ex ante preference" treatment asked which option you think the receiver would prefer the dictator to choose. And the 3rd "moral appropriateness" treatment asked which option you think is more morally appropriate for the dictator to choose. The responses reveal a striking difference between morality and doing what you believe will make other people happy. Equally strikingly, they also reveal a gap between what people are perceived to prefer and what will make them happy.

Consider the choice between sharing the veiled amount, or the veiled amount $+\epsilon$. 22.6% of respondents state that they believe a receiver would be happier after receiving the veiled amount, indicating that knowing about our peers' bad intentions has harmful consequences. But only half as many, 12.4%, state that the receiver would prefer for the dictator to choose the veiled quantity. This raises an interesting question about the relationship between preferences and happiness. Is "ex post overall happiness" imperfectly capturing the concept of total utility, or do people in fact *prefer* overall unhappy outcomes in some circumstances? Is ignorance always bliss or does the act of *choosing* ignorance negate the positive effect on happiness that that it might provide?

The tension between information and emotional utility then leads to difficulty with the concept of morality. Does morality dictate that we should treat other people according to how they would choose, or according to what will make them happiest? *Should* we tell white lies, or should we answer honestly when faced with the proverbial "honey, do these jeans make me look fat?"

Responses from the third phrasing treatment indicate that preserving beliefs-based utility is an important component of morality. The percentage (26.0%) of respondents indicating that sharing the veiled amount is more morally appropriate than the slightly higher amount is on par with the percentage who report that the veiled amount leads to higher ex post happiness (22.6%).

A similar pattern is revealed in the choices between the veiled quantity and a slightly *smaller* amount. Almost no one (2.7%) thinks the smaller amount will make the receiver more happy than the veiled amount, but 10.7% think the receiver would nonetheless prefer that the dictator choose the smaller amount. Fully 15.4% in fact think this is the more morally appropriate option. In light of the 26.0% who think the veiled option is more morally appropriate than the *higher* amount, it seems that there is substantial disagreement on the question of whether it is more morally important to be honest or to protect others' feelings.[11]

The questions about good information also reveal an intriguing pattern. In the "ex ante preference" phrasing treatment, 12.1% of respondents state that receivers would prefer that the dictator share the (very generous) veiled amount rather than a slightly larger share.

[11] Erat and Gneezy (2012) finds a similar tension between norms of truth-telling and norms of other-regarding choices, but others' utility in their experiment is straightforwardly monetary, not beliefs-based.

16.6% state that the receiver would prefer that the dictator share a slightly *smaller* amount than the veiled option.

In terms of ex post happiness, there is a clearer trend. 17.0% believe that good information is worth money to the receiver, while only 6.5% think good information is worth avoiding. But the responses about moral appropriateness are shockingly divergent from either of the other two treatments. Fully 60.0% think it's more morally appropriate to share the veiled amount rather than more, and 50.0% think it's better to share slightly less than the veiled option.

These treatments successfully prove that respondents are not answering questions about ex post happiness as though they were being asked about morality or preferred choices, as intended. But they also provide further evidence that what is good and bad, and therefore how someone with BBA should act, is somewhat ambiguous in this setting. This was a necessary consequence of the design that prioritized identification of BBA separately from social image and outcome-based altruism, but now that the basic phenomenon has been revealed, this will hopefully provide a rich area for further research.

6 Conclusion

Extensive previous evidence for social signaling can also be explained with a more optimistic interpretation of prosocial behavior: If we care about others' subjective utility, independently of their material outcomes, we might avoid hurting others' feelings, or we might seek out ways to express sympathy or solidarity, in ways that can be mistaken for attempts to signal our own altruism. The literature on signaling has almost unilaterally ignored this alternative possibility, and while social signaling is undoubtedly an important driver for such behavior, these results show that a substantial fraction of the population may also be motivated by purer beliefs-based altruism. Up to around 40% of respondents report beliefs consistent with beliefs-based altruism, and these stated beliefs are strongly linked to behavior in an experimental setting. There is additionally no evidence that this link is opportunistic, in the sense that beliefs-based altruism provides an excuse to behave selfishly for the sake of preserving others' blissful ignorance.

The experimental design used in this study is contrived in such a way as to allow clear identification of BBA, but beyond providing a richer understanding of human behavior, these results are merely a first step towards studying BBA in the realworld scenarios in which is more likely to be an important and targetable motivation. These settings, like charitable giving, may not allow for clean identification, but having established the clear existence of BBA for at least a substantial fraction of the population we can now proceed by making likely and simplifying assumptions about the likelihood and relative magnitude of various social image motivations in particular settings. Erat and Gneezy (2012) provides a nice example of how this can work: they find that people are willing to tell altruistic lies (lies that benefit the listener), but are less willing to do so when being watched. As usual, it is possible to explain this using social image concerns exclusively: The listener will judge the liar for being harsh if he doesn't lie, but an outside observer will judge the liar for *failing* to be harshly honest. It's intuitively far more likely that social image pushes in approximately the same direction no matter who the observer is but that the liar has BBA towards the listener's feelings and is thus willing to trade potential social image for them until the social image concerns become too strong.

Another avenue for further research is to explore other forms of beliefs-based preferences and beliefs-based altruism. Behavioral economics has identified many non-material sources of utility in addition to utility from intentions, such as anticipatory utility (based on beliefs about future events), ego utility (based on beliefs about one's own abilities), loss aversion (based on information about a status quo or expected outcome), and many others. It's therefore natural to suppose that people's known altruistic preferences might extend to these nonmaterial sources of utility. BBA could therefore be more broadly defined as altruism over any non-material aspect of utility. Reference-dependent social preferences are one type of BBA, according to this definition, that has gotten a small amount of attention (Breitmoser and Tan, 2014, 2013). Another type of BBA, altruism over feelings of solidarity or likemindedness is also used by Rotemberg (2009) in a model of voting. Yet another type, altruism over feelings of regret, is used to explain why firms fail to price gouge opportunistically in Rotemberg (2011). Rotemberg (2014) provides a nice review of other studies that are consistent with other types of BBA, but like the signaling results mentioned in the introduction, these results are universally explainable with social signaling, and none systematically attempt to disentangle the two. Battigalli and Dufwenberg (2007) and Charness and Dufwenberg (2006) study "simple guilt" aversion, which is the desire to prevent others from being disappointed (relative to expectations). This is in fact a concept closely related to BBA, as described in footnote 3. Overall, moving from a concept of altruism tied to monetary/consumption-based outcomes to a more general concept that recognizes all sources of utility promises to be a fruitful approach.

References

References

Alpizar, Francisco, Fredrik Carlsson, and Olof Johansson-Stenman. 2008. "Does context matter more for hypothetical thanforactual contributions? Evidence from a natural field experiment." *Experimental Economics*, 11(3): 299–314.

Amir, Ofra, David G. Rand, and Ya'akov Kobi Gal. 2012. "Economic games on the internet: The effect of $1 stakes." *PLoS ONE*, 7(2): e31461.

Andreoni, James, and Alison Sanchez. 2014. "Do beliefs justify actions or do actions justify beliefs? An experiment on stated beliefs, revealed beliefs, and social-image manipulation." NBER Working Paper Series No. 20649.

Andreoni, James, and B. Douglas Bernheim. 2009. "Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects." *Econometrica*, 77(5): 1607–1636.

Andreoni, James, and Ragan Petrie. 2004. "Public goods experiments without confidentiality: a glimpse into fund-raising." *Journal of Public Economics*, 88(7-8): 1605–1623.

Ashraf, Nava, Oriana Bandiera, and Kelsey Jack. 2014. "No margin, no mission? A field experiment on incentives for pro-social tasks." *Journal of Public Economics*, 120: 1–17.

Babcock, Philip, Kelly Bedard, Gary B. Charness, John Hartman, and Heather Royer. 2015. "Letting down the team? Evidence of social effects of team incentives." *Journal of the European Economic Association*, 13(5): 841–870.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2005. "Social preferences and the response to incentives: Evidence from personnel data." *Quarterly Journal of Economics*, 120(3): 917–962.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2007. "Guilt in games." *American Economic Review*, 97(2): 170–176.

Bohannon, John. 2011. "Social science for pennies." *Science*, 334(October): 307.

Bohnet, Iris, and Bruno S. Frey. 1999. “The sound of silence in prisoner’s dilemma and dictator games.” *Journal of Economic Behavior & Organization*, 38(1): 43–57.

Breitmoser, Yves, and Jonathan H.W. Tan. 2013. “Reference dependent altruism in demand bargaining.” *Journal of Economic Behavior and Organization*, 92: 127–140.

Breitmoser, Yves, and Jonathan H.W. Tan. 2014. “Reference dependent altruism.” Münich Personal RePEc Archive Paper No. 52774.

Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson. 2007. “Is generosity involuntary?” *Economics Letters*, 94(1): 32–37.

Cason, Timothy N., and Feisal U. Khan. 1999. “A laboratory study of voluntary public goods provision with imperfect monitoring and communication.” *Journal of Development Economics*, 58(2): 533–552.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci. 2014. “Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers.” *Behavior Research Methods*, 46: 112–130.

Charness, Gary B., and Martin Dufwenberg. 2006. “Promises and partnership.” *Econometrica*, 74(6): 1579–1601.

Dana, Jason, Daylian M. Cain, and Robyn M. Dawes. 2006. “What you don’t know won’t hurt me: Costly (but quiet) exit in dictator games.” *Organizational Behavior and Human Decision Processes*, 100: 193–201.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. “Testing for altruism and social pressure in charitable giving.” *Quarterly Journal of Economics*, 127(1): 1–56.

Dellavigna, Stefano, John A List, Ulrike Malmendier, and Gautam Rao. 2017. “Voting to Tell Others.” *Review of Economic Studies*, 84(October 2016): 143–181. NBER Working Paper Series No. 19832.

Di Tella, Rafael, Ricardo Pérez-Truglia, Andres Babino, and Mariano Sigman. 2015. “Conveniently upset: Avoiding altruism by distorting beliefs about others.” *American Economic Review*, 105(11): 3416–3442.

Erat, Sanjiv, and Uri Gneezy. 2012. “White lies.” *Management Science*, 58(4): 723–733.

Franzen, Axel, and Sonja Pointner. 2012. "Anonymity in the dictator game revisited." *Journal of Economic Behavior & Organization*, 81(1): 74–81.

Gerber, Alan S., Donald Philip Green, and Christopher W. Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American Political Science Review*, 102(01): 33–48.

Grossman, Zachary. 2008. "Signaling, beliefs, and prosocial behavior." PhD diss. University of California at Berkeley.

Grossman, Zachary. 2015. "Self-signaling and social-signaling in giving." *Journal of Economic Behavior & Organization*, 117: 26–39.

Hoffman, Elizabeth, Kevin A. McCabe, Keith Shachat, and Vernon L. Smith. 1994. "Preferences, property rights, and anonymity in bargaining games." *Games and Economic Behavior*, 7(3): 346–380.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. "The online laboratory: Conducting experiments in a real labor market." *Experimental Economics*, 14(3): 399–425.

Ipeirotis, Panagiotis G. 2010. "Demographics of Mechanical Turk." NYU Stern School of Business CeDER Working Papers No. 10-01.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber. 2012. "Sorting in experiments with application to social preferences." *American Economic Journal: Applied Economics*, 4(1): 136–164.

Levine, David K. 1998. "Modeling altruism and spitefulness in experiments." *Review of Economic Dynamics*, 1(3): 593–622.

Malmendier, Ulrike, Vera L. te Velde, and Roberto A. Weber. 2014. "Rethinking reciprocity." *Annual Review of Economics*, 6: 849–874.

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at work." *American Economic Review*, 99(1): 112–145.

Mason, Winter A., and Siddharth Suri. 2010. "Conducting behavioral research on Amazon's Mechanical Turk." *Judgment and Decision Making*, 5(5): 411–419.

Paolacci, Gabriele, Katherine A. Burson, and Scott I. Rick. 2011. "The intermediate alternative effect: Considering a small tradeoff increases subsequent willingness to make large tradeoffs." *Journal of Consumer Psychology*, 21: 384–392.

Rabin, Matthew. 1993. "Incorporating fairness into game theory and economics." *American Economic Review*, 83(5): 1281–1302.

Rege, Mari, and Kjetil Telle. 2004. "The impact of social approval and framing on cooperation in public good situations." *Journal of Public Economics*, 88(7-8): 1625–1644.

Rotemberg, Julio J. 2009. "Attitude-dependent altruism, turnout and voting." *Public Choice*, 140(1/2): 223–244.

Rotemberg, Julio J. 2011. "Fair pricing." *Journal of the European Economic Association*, 9(5): 952–981.

Rotemberg, Julio J. 2014. "Models of caring, or acting as if one cared, about the welfare of others." *Annual Review of Economics*, 6(6): 1–26.

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

te Velde, Vera L. 2016. "Large-sample subject recruitment and selection on Amazon's Mechanical Turk." Working Paper.

A Example game instructions

Page 1 of 3

Welcome!

My name is Vera te Velde, and I am a researcher at the University of California at Berkeley. This short set of questions will help us understand how you think about certain economic and social situations.

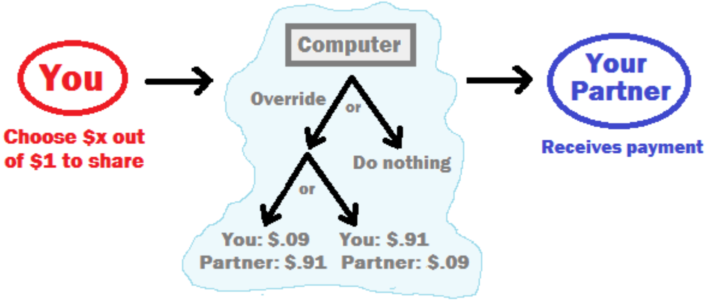
There are no right or wrong answers, but you must answer carefully and honestly in order to avoid directly contradicting yourself (for example, by saying you prefer an apple to an orange, but you'd rather have an orange than *two* apples). As long as you answer honestly, you don't need to worry about making that kind of mistake, and you will be eligible for bonus payments!

Game

First, you will have the opportunity to play a simple game for real money with a real partner. You have been matched with another MTurk user. You will both play the exact same game, shown below. Either your choice or your partner's choice will be randomly selected, and then with a 20% probability, the game will count for real money and you will be paid according to the results with an MTurk bonus payment. The results of the game will be reported to you on MTurk after both you and your partner have completed the HIT, along with your bonus payment (if applicable).

In this game, you must decide how to divide $1 between yourself and your partner. After you decide, the computer will draw a random number, and half the time it will override your choice. If it does, half the time it will give 9 cents to you and 91 cents to your partner, and half the time it will choose to give 91 cents to you and 9 cents to your partner. Your partner will see the outcome of the game, but will *not* find out whether the computer overrode your choice.

Overall, 25% of the time the computer will share 9 cents with your partner, 25% of the time it will share 91 cents, and half the time it will share whatever amount you choose below.



1. How much would you like to share with your partner? Choose an amount between $0 and $1: $1 ▼

2. What do you think most other people will answer to the previous question? For this question, if you guess within 10 cents of the middle answer among MTurk users who complete this HIT, you will receive a 5 cent bonus payment! $1 ▼

If you have any thoughts or comments, or if you would like to share your reasoning, please do so here. We really appreciate your input!

Continue

This is a study by the University of California, Berkeley, Department of Economics. For questions, contact Vera te Velde at vtevelde@econ.berkeley.edu.

B Example questionnaire instructions

Page 2 of 3

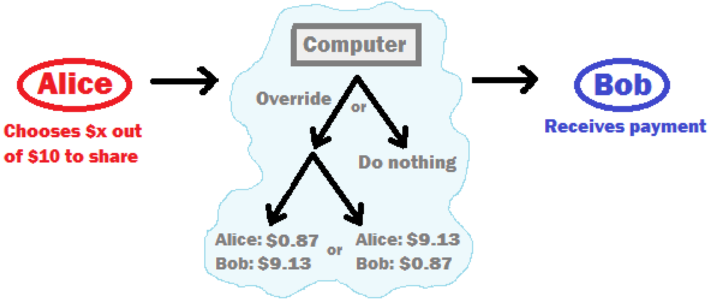
Questionnaire

Next I will ask you a few questions about a game very similar to the one you just played. Imagine the following scenario. Two people, Alice and Bob, are playing a game through a computer. They don't know each other and will never actually meet, and they are playing this game together just one time.

In this game, Alice must decide how to divide $10 between herself and Bob. After she decides, half the time they each receive their money and the game ends. The other half of the time, the computer randomly decides to override Alice's choice. Half the time the computer decides to give Alice $9.13 and Bob $0.87, and half the time it gives Alice $0.87 and Bob $9.13.

Alice and Bob know the rules of the game, and Bob can see how much money he receives, *but he doesn't know whether the computer actually intervenes or what it randomly chooses.*

Overall, 25% of the time the computer shares $0.87 with Bob, 25% of the time it shares $9.13 with Bob, and half the time it shares whatever quantity Alice chooses.



Notice: If Alice chooses to give Bob either $0.87 or $9.13, Bob cannot figure out whether Alice actually chose to share that amount or whether the computer overrode her choice. If Alice chooses to give Bob any other amount, he can figure out what she chose.

The following few questions ask you to guess or express an opinion about some of the possible outcomes in this game. Imagine that Bob and Alice are average people and answer what you think they would do or think.

1. How much do you think Alice will choose to share with Bob? Enter an amount between $0.00 and $10.00.

□

2. What do you think most *other* people will answer to the previous question? For this question, if you guess within 25 cents of the middle answer among other MTurk users who complete this HIT, you will receive a 5 cent bonus payment!

□

3. Do you think Bob would be happier after receiving a payment of $0.87 or $0.92?

○ $0.87
○ $0.92

4. Do you think Bob would be happier after receiving a payment of $0.87 or $0.82?

○ $0.82
○ $0.87

5. Do you think Bob would be happier after receiving a payment of $9.13 or $9.18?

○ $9.13
○ $9.18

6. Do you think Bob would be happier after receiving a payment of $0.87 or $0.92?

○ $9.13
○ $9.08

7. Do you think Bob would be happier after receiving a payment of $0.87 or $1.27?

○ $1.27
○ $0.87

8. Do you think Bob would be happier after receiving a payment of $0.87 or $0.47?

○ $0.47
○ $0.87

9. Do you think Bob would be happier after receiving a payment of $9.13 or $9.53?

○ $9.13
○ $9.53

10. Do you think Bob would be happier after receiving a payment of $9.13 or $8.73?

○ $8.73

○ $9.13

If you have any thoughts or comments, or if you would like to share your reasoning, please do so here. We really appreciate your input!

Next

This is a study by the University of California, Berkeley, Department of Economics. For questions, contact Vera te Velde at vtevelde@econ.berkeley.edu.

C Demographic survey

Page 3 of 3

Survey

Please answer the following five questions to complete the study:

What is your age bracket?

○ Under 30

○ 30-39

○ 40-49

○ 50-59

○ 60 or over

○ I prefer not to answer

Are you male or female?

○ Female

○ Male

○ Other, or I prefer not to answer

What is your ethnicity?

○ White

○ Non-white hispanic

○ Black or African American

○ Asian or Pacific Islander
○ Other
○ I prefer not to answer

What is your household income?
○ Less than $20,000
○ $20,000 to $49,999
○ $50,000 to $79,999
○ $80,000 to $149,999
○ $150,000 or more
○ I prefer not to answer.

What is your highest completed education level?
○ High school or less
○ Some college
○ Bachelor’s degree
○ Postgraduate degree
○ I prefer not to answer.

[Finish]

This is a study by the University of California, Berkeley, Department of Economics. For questions, contact Vera te Velde at vtevelde@econ.berkeley.edu.