Förster, Manuel; van der Weele, Joël J.

Working Paper

Persuasion, justification and the communication of social impact

Tinbergen Institute Discussion Paper, No. TI 2018-067/I

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Förster, Manuel; van der Weele, Joël J. (2018) : Persuasion, justification and the communication of social impact, Tinbergen Institute Discussion Paper, No. TI 2018-067/I, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/185586

TI 2018-067/I
Tinbergen Institute Discussion Paper

Persuasion, justification and the communication of social impact

Manuel Foerster[1]
Joel (J.J.) van der Weele[2]

[1] University of Hamburg
[2] Universiteit van Amsterdam

Persuasion, justification and the communication of social impact

Manuel Foerster* Joël J. van der Weele◇

August 25, 2018

Abstract

We experimentally investigate strategic communication about the impact of prosocial actions, which is central to policy debates about foreign aid or the environment. In our experiment, a "sender" receives an informative but noisy signal about the impact of a charitable donation. She then sends a message to a "receiver", upon which both subjects choose whether to donate. The sender faces a trade-off between persuading the receiver to act and justifying her own inaction. We find evidence for both motives. Increasing the visibility of the sender's actions increases the justification motive and makes senders more likely to report low impact, reducing giving among receivers. These results show the intimate links between reputation and communication in moral domains, and help understand the fraught nature of political discussions about social impact.

JEL classification: C91, D83, D91

Keywords: cheap talk, image concerns, information aggregation, charitable giving, economic experiments.

*University of Hamburg, email: manuel.foerster@uni-hamburg.de.

◇University of Amsterdam, Tinbergen Institute, email: vdweele@uva.nl.

We would like to thank Christine Exley, Ivan Soraperra, Jeroen van de Ven, Jason Dana, Peter Schwardmann and Florian Zimmermann for useful comments. We thank Alejandro Miranda Salas, Davide Pace and Ivar Kolvoort for excellent research assistance. Joël van der Weele gratefully acknowledges financial support from the NWO through VIDI grant 452-17-004. We report all data gathered in the context of this research project.

1 Introduction

Parties in policy debates often agree on ethical principles but disagree on the right policy to implement them. For instance, both proponents and opponents of development aid agree that helping people in need is morally right, but disagree about the effectiveness of aid. Proponents argue that aid saves life and encourages development, while opponents maintain that it mainly feeds corruption.[1] Surveys reflect these different viewpoints. The Eurobarometer (2005) survey shows that 91 percent of European citizens agree that helping people in poor countries develop is important, but only about half say that aid is effective in doing so. In the UK, around 57 percent of the public reports that "much development assistance is wasted and does little or nothing to promote British interests and should therefore be radically reduced" (Chatham House and YouGov, 2011, see also Bond, 2015). In a recent, large scale study about charitable motives among Canadians, 61 percent said they would give more if they had more confidence in charities and where the money is going (Angus Reid Institute, 2017).

Regardless of their factual basis, public expressions of skepticism or enthusiasm of foreign aid may reflect strategic considerations. Theoretical arguments by Foerster and van der Weele (2018, FvW18 hereafter) and Bénabou et al. (2018) highlight a central trade-off in the communication about ethical actions. Publicly communicating a high social impact may persuade others to contribute and increase social welfare. At the same time, it raises moral pressure on the communicator to personally take costly action. A wish to escape that pressure may lead one to downplay the social returns as a justification for inaction. The balance between these motives determines political communication and public opinion about the impact of foreign aid, climate change or other actions with moral implications.

Empirically, little is known about how strategic motives affect communication about social impact, and how this affects personal behavior and the common good. To understand the balance between persuasion and justification, we conduct a laboratory experiment on charitable giving. An informed "sender" is matched with an uninformed "receiver", both of whom may choose to make a donation to a charity, GiveDirectly. Before they do so, the sender receives a noisy but informative signal about the impact of the donation. She can then communicate a message to the receiver, with the option to falsify the signal she observed. In a "public" treatment, we make the actions of the sender visible and salient to other participants, thus creating a justification motive for unwilling givers.

Even though roughly half of the participants always communicate honestly, we find evidence for a persuasion motive, as more than 40 percent of the subjects in the private

[1] For instance, the British politician Nigel Farage has argued to abandon what he calls "the Foreign Bribery Budget", and U.S. senator Jesse Helms has compared foreign aid with throwing tax dollars down a "Foreign Rat Hole". For Farage, see `https://www.lbc.co.uk/radio/presenters/nigel-farage/nigel-lets-call-it-the-foreign-bribery-budget/`, accessed April 22, 2018. For Helms, see `https://www.washingtontimes.com/news/2001/jan/12/20010112-020658-4661r/`, accessed April 22, 2018.

treatment exaggerate impact at least sometimes. This happens despite a lack of personal benefits from the other's contributions. Moreover, in line with a motive to justify selfish actions, senders in the public treatment are significantly less likely to report high impact. Image concerns appear to drive this difference; in a bid to avoid looking hypocritical, senders in the public treatment are more likely to donate after a high message, thus making this message more costly. The shift in communication has a large impact on the behavior of receivers, who are about 40 percentage points less likely to donate after receiving a low signal.

To our knowledge, we are the first to empirically investigate the motives of persuasion and justification in the communication of social impact. Our main insight is that persuasion motives lead to meaningful exaggeration, while image concerns lead to downplaying of social impact. This is an important addition to the literature on audience effects that has mainly emphasized their positive impact on prosocial behavior.[2] More generally, the experiment demonstrates how communication about externalities is intimately tied to public image management. It helps explain why efficient information aggregation about issues like foreign aid, climate change mitigation or other public goods cannot be taken for granted in the political arena, where image and reputation are of paramount importance.

Previous investigations of the justification motive have focused on the use of "moral wiggle room". People look to sidestep image concerns or social pressure to give by exploiting plausible deniability for their selfish actions, like uncertainty about impact (Dana et al., 2007; Andreoni and Bernheim, 2009). If given the choice, a substantial proportion of subjects in the laboratory chooses not to know information about the impact, even if they would use the information when they are forced to observe it (Dana et al., 2007; Grossman and van der Weele, 2017). Exley (2016a,b) shows that people self-servingly interpret information about the risky impact and effectiveness of the donation. Unlike much of the moral wiggle room literature, we focus on the willingness to (consciously) misrepresent information, and the resulting impact on others' decisions. This also sets our paper apart from Andreoni and Sanchez (2014), who show that subjects in the lab falsify belief statements to the experimenter to justify their actions, but such falsification has no further impact on other people's decisions.

This focus on misrepresentation of information relates our paper to a literature on lying and deception. Many studies show that people lie for money, although not everyone does so (e.g. Gneezy, 2005; Abeler et al., 2018). We show that about half of the participants are willing to misrepresent information without *any* pecuniary benefit to themselves. Although misrepresentation is not technically the same as lying, it shows that we should take into account image and persuasion motives alongside monetary ones

[2] In line with theories of costly signaling of altruism, increasing the visibility of donations raises prosocial behavior in various contexts, as has been demonstrated in the lab (e.g. Rege and Telle, 2004; Andreoni and Petrie, 2004; Ariely et al., 2009) and in the field (e.g. Harbaugh, 1998; Soetevent, 2005; Lacetera and Macis, 2010; Karlan and McConnell, 2014).

when studying communication.

Previous studies have also documented the importance of social impact. In addition to the surveys mentioned above, the experimental literature indicates that information about effectiveness (Gordon et al., 2009; Metzger and Günther, 2015; Karlan and Wood, 2017; Yörük, 2016) or assurances that the donations will not be used for overhead (Gneezy et al., 2014) matter for charitable giving. In a laboratory experiment, Butera and Horn (2017) vary both information about effectiveness and public visibility of donations. They find that information about effectiveness reduces giving under public visibility, in line with a donor motive to signal concern for giving effectively. Here, we show that strategic misrepresentation about social impact arises exactly because it is such an important determinant of prosocial decisions.

On the theoretical side, the experiment informs the formal models by FvW18 and Bénabou et al. (2018) that we will discuss in more detail in the hypothesis section. The results are also related to the seminal and influential work by Kuran (1987, 1997), who argues that social pressures lead to "preference falsification", i.e. strategic misrepresentation of their values or information.

2 Experimental design

Sessions were run at the CREED laboratory at the University of Amsterdam, lasting about one hour each.[3] Subjects were recruited using the online CREED recruitment system. In total, 228 subjects participated in 14 sessions: 7 sessions with a total of 116 subjects in the *Public* treatment and 7 sessions with a total of 112 subjects in the *Private* treatment. Each session had between 12 and 18 participants, always in even numbers, depending on the show-up rate.

The charity in our experiment was GiveDirectly. GiveDirectly makes direct cash transfers to poor recipients in East Africa, and 91 percent of each donation ends up with the recipient. This charity is convenient as its activities are easy to explain.[4] Subjects were given assurances that each donation would actually be transferred on their behalf by the experimenter, referring to the no-deception policy of the CREED laboratory. All instructions are available in Appendix B.[5]

2.1 Timing and procedures

Upon entering the lab, participants were randomly allocated a seat at a computer terminal. The experimenter read aloud the instructions for the first part of the experiment,

[3] 14 sessions of the current experiment were run in February 2018. We had to discard the data of one session in which a technical problem occurred, so we ran one additional session in July 2018.

[4] See `www.givedirectly.org` for more details of the charity's activities.

[5] The IRB of the University of Amsterdam approved the experiment. We preregistered the experiment at `www.aspredicted.org`. Both IRB approval and preregistration are available on request. While the experiment was designed to be able to look at both under- and overreporting, our preregistration focused on underreporting which we expected to be more prevalent.

which contained information about the activities of GiveDirectly and about a social value orientation (SVO) task with GiveDirectly as recipient. The SVO measures willingness to contribute to the charity, and consists of six consecutive allocation tasks between the participants and the charity, where the possible allocations change in each decision. We followed the slider-based SVO design by Murphy et al. (2011) and used the program by Crosetto et al. (2012), programmed in the software z-tree (Fischbacher, 2007).

After the SVO task was completed, the experimenter distributed the instructions for the main part of the experiment, the interaction stage. After reading these instructions, participants answered a few control questions to test their understanding of the payment scheme. They then learned their role as sender or receiver, and moved to the interaction tables in the lab. Senders and receivers were seated opposite each other. We chose face-to-face interactions in order to strengthen image concerns in the *Public* treatment.[6] Between each sender-receiver pair there was a divider, so participants had no contact with the adjacent pair. Subjects were told that verbal communication was not allowed and would result in exclusion from payment in the experiment. No participant was caught communicating verbally or non-verbally in other ways than described in the instructions.

Before the first interaction round, participants completed a practice round to familiarize themselves with the procedures. Each interaction round proceeded as follows. The experimenter privately rolled a die to determine the "type of the interaction" for that round, either "red" or "green" with equal probability. Participants did not know the type of the interaction, but senders were given a noisy signal. They were asked to draw one card from a deck, which consisted of two red cards and one green card if the true type was red, and two green cards and one red card if the true type was green. Thus, the signal was correct with a probability of two thirds. Upon receiving the signal, the sender communicated with the receiver by showing either a red or a green card. Finally, both sender and receiver chose between two options described in detail below.

Both subjects recorded all choices on a decision sheet that was placed behind a screen on the table and only visible to them. To ensure truthful reporting of the communicated card, subjects were told they would only be paid if both members of the pair recorded the same color. After each interaction round receivers left their seat and moved one place to the left. At the end of the interaction period, one of the rounds was randomly drawn for payment. All participants then returned to their cubicle and answered a short questionnaire. Finally, participants were paid their earnings privately and in cash.

Payment for participants consisted by summing a show-up payment of €6, one randomly drawn choice (from six) in the SVO task (paying between €0.50 and €1) and the earnings from one randomly drawn round in the interaction stage (paying between

[6]To further raise social pressure to give, each table featured a sheet with a testimonial and photo of a potential recipient, taken from the website of GiveDirectly (see Appendix B for an example). At the beginning of each round, participants were invited to read the testimonial. While senders remained seated, receivers changed table and would thus be confronted with a new testimonial each time.

€0 and €10, with payment based on the decision sheets). The average subject earned €14.40 (min. €6.50, max. €17). The money generated for the charity in the selected round was summed up after the experiment and transferred by the experimenter.

2.2 Payoffs in the interaction stage

At the end of each interaction round, each subject had to choose between *Option 1* and *Option 2*. As depicted in Figure 1, the payoff of *Option 1* depended on the type of the interaction. If the type was red, neither the subject nor the charity would earn anything. If the type was green, the charity would earn €15 and the participant €5. The payoff from *Option 2* was independent of the type of interaction and yielded €10 for the participant and €0 for the charity. Below, we sometimes refer to *Option 1* as a "donation", since the participant gives up at least €5 to potentially donate €15 to the charity. The payoffs are depicted in Figure 1, which is taken from the experimental instructions.

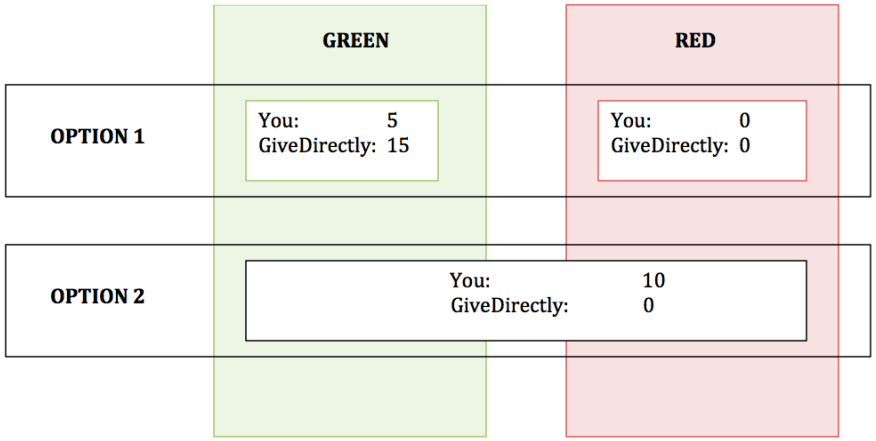


Figure 1: Payoffs in the interaction phase of the experiment.

2.3 Treatments

The experiment has two treatments administered on the session level, i.e. between subjects. In the *Private* treatment, participants' choices between *Option 1* and *Option 2* were private and not visible to any other participant. The *Public* treatment featured two procedures that were designed to make image concerns as strong and salient as possible for the sender. First, at the end of each round, senders would communicate their choice to the receiver in front of them. Second, at the end of the experiment, senders would stand up to announce to all participants in the session their choices in the round that was randomly drawn for payment. In particular, they would publicly announce (a) their communication choice (green/red) and (b) their choice between *Option 1* and *2*.

This procedure was announced in the instructions and applied to the practice round of the experiment.

3 Hypotheses

Our hypotheses are based on FvW18, who develop a cheap-talk model of communication in environments with externalities. Agents can engage in a prosocial action with an individual cost and an unknown social return. Before choosing whether to act or not, agents receive a noisy but informative signal about the impact of the action and then submit a cheap-talk report about the signal to the other agent. Our experimental design closely tracks the application of their model to charitable giving in Section 7. Specifically, the model and experiment have identical timing and use binary signals and choices for communication and actions.[7] We do not reproduce the model here in mathematical detail, as the hypotheses are relatively intuitive. Independent and contemporaneous work by Bénabou et al. (2018) delivers similar intuitions as FvW18, although the experimental design is somewhat further removed from their model.[8]

Agents in FvW18 have different levels of "intrinsic motivation" to be prosocial, modeled as an individual, psychological payoff proportional to the social benefit of the action. The size of this payoff is private information and is distributed according to a common prior. Agents also have "image concerns", a benefit of being thought of by others as a "good person", i.e. someone with a high intrinsic motivation for the prosocial action. FvW18 show that a trade-off between persuasion and justification arises for senders. While reporting a high impact might persuade others to contribute, it also raises pressure on the sender to take the costly prosocial action herself. Not contributing after reporting a high signal means you are seen as a "hypocrite" and get a low image. By contrast, reporting a low impact justifies inaction, but might discourage contributions by others.

FvW18 show that the resolution of the trade-off depends on the relative level of image concerns and intrinsic motivation of the agent. If image concerns are low relative to intrinsic motivation to contribute and agents benefit (psychologically) from others' contributions, then agents report all signals as high signals to persuade others to contribute, referred to as "alarmism" in FvW18's terminology. If image concerns are high enough, then the theory predicts that types with high intrinsic motivation contribute after receiving and reporting a high private signal to avoid being seen as a "hypocrite". They therefore are not in need of justification and report both high and low signals truthfully.

[7] The main difference is that FvW18 consider a continuous state and symmetric players who both act as sender and receiver.

[8] The main differences from Bénabou et al. (2018) is that they give a central role to "narratives". With some imagination, our paper may be seen as giving an extremely stylized operationalization of the use of narratives, but certainly more work remains to be done here. Another important aspect of Bénabou et al. (2018) that is not present in this research is the idea that people may strategically search for narratives. Again, investigation of this idea remains an important task for future research.

Types with low intrinsic motivation are never motivated to contribute. Therefore, when they get a high signal, they misreport and communicate a low signal to justify their inaction. This allows them to pool with a high motivation type who received a low signal and obtain a better image in equilibrium, even though this action may rob the charity of a donation that the receiver might otherwise have made. Overall, equilibrium communication will feature partial downplaying of the social return (or "denial" in FvW18's terminology).[9]

To turn these insights into hypotheses for our experiment, we introduce some terminology to simplify the description of communication strategies. First, we define "honesty" as the strategy to communicate the signal accurately. Second, "playing down" or "underreporting" refers to the strategy to report a red card after seeing a green card. Third, "persuading" or "overreporting" refers to the strategy to report a green card after seeing a red card.[10]

As a point of departure, we are interested in the absolute levels of honesty, under- and overreporting. The theory does not generate detailed hypotheses on these absolute levels, but we can make some informed remarks. First, we know from previous literature that many people are honest even if this is not in their monetary interest. In this experiment, there is no monetary incentive for misreporting, so we would expect substantial amounts of honesty as well. Second, the theory in FvW18 predicts that persuasion will occur if image concerns are low (i.e. in the *Private* treatment) and agents benefit psychologically or monetarily from others' contributions. Since there is no monetary benefit from the other's contribution in this experiment, persuasion and overreporting will be driven entirely by non-monetary motivation.[11] Third, in the *Private* treatment there is no justification motive, so we do not expect (much) underreporting in this condition.

The theory yields an unambiguous prediction for the treatment manipulation. Raising image concerns in the *Public* treatment will increase the justification motive, making the report of a green card more costly.

Hypothesis 1 (Justification motive)**.** *Senders will show fewer green cards in the Public treatment.*

The theory predicts two further behavioral patterns stemming from the justification motive. First, justification arises because there is a "price" on the communication of

[9] As is common in cheap-talk settings, there are also "babbling" equilibria in which no information is transmitted. We focus here on the unique equilibrium with "influential communication" as defined in FvW18.

[10] Notice that different to FvW18, we do not define "underreporting" ("overreporting") as the strategy in which the agent reports a red (green) card regardless of her signal. In each round of the experiment, we can only distinguish playing down (exaggerating) from honesty in case the sender has seen a green (red) card and reported a red (green) card. We therefore treat instances in which the report matches the card seen as honesty.

[11] We ran a few pilot sessions of an alternative design that included monetary spillovers between participants. This design appeared too complicated for participants to understand (more details are available on request). Including monetary spillovers, as in a public good setting, therefore remains an interesting area for future research.

a green card: the sender is now supposed to be prosocial. Not doing so is considered "hypocritical", and leads to a low image.

Hypothesis 2 ("Price" on high report)**.** *Senders are more likely to choose Option 1 after showing a green card in the Public treatment.*

Second, the theory predicts that underreporting is a strategy used in the *Public* treatment, but only by senders who have low motivation to contribute. In the experiment, we measure the motivation for "sorting" into communication strategies by our SVO task.

Hypothesis 3 (Sorting)**.** *In the* Public *treatment, senders with a low SVO score are more likely to underreport.*

Our last hypothesis relates to the impact of communication on the receivers. The model predicts that subjects should react to a high signal by contributing, at least those who are motivated to contribute. This is because in equilibrium, high messages are informative of a high social return.

Hypothesis 4 (Communication impact)**.** *Receivers are more likely to choose Option 1 after seeing a green card.*

4 Results

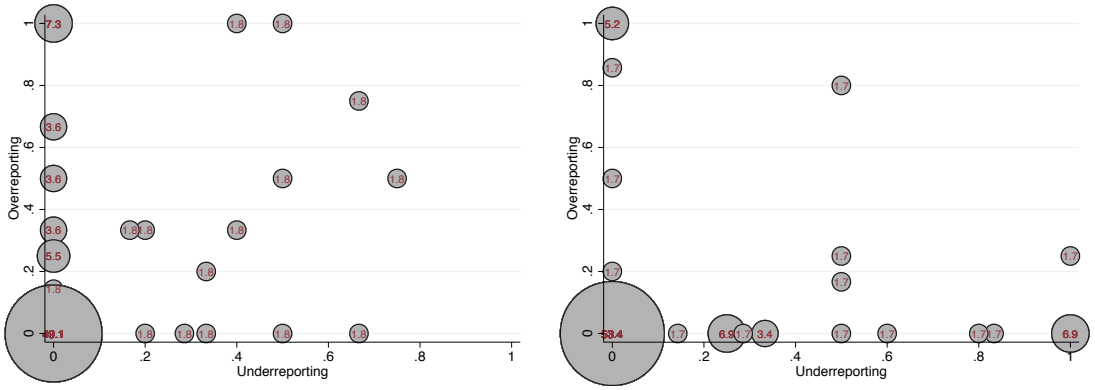
We start our analysis by providing a descriptive overview of senders' communication.[12] Table 1 shows descriptive statistics of the fraction of green cards shown split by treatment and by the color of the card drawn by the sender. It is immediately clear that senders overwhelmingly communicate more green cards after drawing a green card, pointing at high levels of honesty. Furthermore, more green cards are shown in the *Private* treatment, regardless of which card was drawn.

	Private	Public	Total
Green card drawn	0.90	0.79	0.84
	(0.29)	(0.41)	(0.36)
Red card drawn	0.22	0.13	0.17
	(0.41)	(0.33)	(0.38)
Total	0.56	0.49	0.52
	(0.50)	(0.50)	(0.50)

Table 1: **Green cards shown by senders as a fraction of all cards shown**, split by treatment and the color of card drawn. Standard deviation in parenthesis.

[12] We excluded observations from one participant who did not follow the instructions. The participant could not answer the questions to check understanding without extensive help. In the role of sender, the subject drew multiple cards from the deck more than once, invalidating the signal. In addition, we excluded data from two interactions where the sender's and receiver's report of the communicated message did not match, so we do not know which message was communicated.

The fractions in Table 1 do not take into account the dependence of observations that belong to the same sender. Therefore, Figure 2 provides a graphical overview of the communication patterns in the two treatments, where we represent each participant as an individual data point. On the horizontal axis we plot the fraction of green cards that are misrepresented as red cards (the underreporting or "justification axis"), and on the vertical axis the fraction of red cards misrepresented as green cards (the overreporting or "persuasion axis"). Thus, an honest individual is located at the origin (or the South-West corner), someone who always overreports and never underreports is in the North-West corner, whereas someone who always underreports and never overreports is in the South-East corner. The size of the observation marker reflects the percentage of total observations in the treatment, which is stated inside the marker.



(a) *Private* treatment (b) *Public* treatment

Figure 2: **Overview of communication strategies.** The x-axis shows the degree of underreporting: the fraction of green cards that are misreported as a red card. The y-axis shows the degree of overreporting: the fraction of red cards that are misreported as a green card. A person who is honest in all rounds is located at the origin. The size of the observation marker reflects the percentage of total observations in the treatment, which is stated inside the marker.

It is immediately clear from the figure that about half of the participants in both treatments is always honest. Of the remaining half, there is a shift from overreporting to underreporting between the *Private* and *Public* treatment. In the *Private* treatment 42 percent overreport at least sometimes, and 11 percent always do so, whereas in the *Public* treatment these numbers are 18 percent and 5 percent, respectively. Overreporting in the *Private* treatment is substantial, and suggests the presence of a persuasion motive despite a lack of personal benefits from others' contributions. We discuss this possibility more extensively in Section 4.2. When it comes to underreporting, 25 percent do so at least sometimes and no subject always does so in the *Private* treatment, while in the *Public* treatment these numbers are 36 percent and 9 percent, respectively.

There are several ways to statistically compare the two-dimensional distributions in Figure 2. The most straightforward way is to test Hypothesis 1 and ask how many green

cards senders reported in each treatment, shown in Table 1. For an adequate test, we should control for the amount of green cards drawn in each treatment, as there turned out to be a slightly (but insignificantly) higher proportion in the *Public* treatment.

	(1) Linear RE	(2) Linear RE	(3) Probit MFX	(4) Probit MFX
Public treatment (PT)	-0.103***	-0.114**	-0.101***	-0.136***
	(0.0399)	(0.0455)	(0.0380)	(0.0458)
Green card drawn (GC)	0.680***	0.670***	0.485***	0.449***
	(0.0218)	(0.0313)	(0.00307)	(0.0257)
PT x GC		0.0200		0.0693
		(0.0436)		(0.0490)
Constant	0.217***	0.223***		
	(0.0307)	(0.0325)		
Observations	931	931	931	931

Table 2: **Regressions of green card shown on treatment and card drawn.** Columns (1) and (2) show a linear model, columns (3) and (4) show random effects probit marginal effect estimations. Both models include individual random effects. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$

Table 2 shows the result of such a multivariate analysis. Columns (1) and (2) report the result of a linear probability model (OLS), whereas columns (3) and (4) report the marginal effects of a Probit model. In both instances, the model includes random effects for the sender's ID to correct for dependence between an individual sender's choices. The results from columns (1) and (3) show that drawing a green signal makes a sender between 48 and 67 percent more likely to show a green card, which is compatible with the observation that senders are honest most of the time. In line with Hypothesis 1 and a justification motive, the *Public* treatment reduces the likelihood of showing a green card by about 10 percent, a result that is significant at the 1 percent level on a two-sided test in both specifications.[13] Note that unless otherwise stated, all our statistical tests are two-sided.

Columns (2) and (4) add an interaction term for the treatment and the color of the card drawn by the sender. This shows that the decline in green cards shown is smaller after the sender drew a green card, although the coefficient is not statistically significant in either specification. We perform a χ^2 test for the significance of the sum of the coefficients of the treatment dummy and interaction term, to directly evaluate the effect of the treatment after drawing a green card. We find marginal significance ($p = 0.035$) of the sum of coefficients in column (2), but no significance ($p = 0.14$) in column (4). In Appendix A, we provide a more detailed breakdown of the statistical results separated by the color of the card drawn by the sender.

A more conservative statistical approach is to compare both reporting distributions

[13] These results are robust to clustering standard errors on either ID or session level: we find significance at the 1 percent or 5 percent level, depending on the model.

non-parametrically. To do so, we summarize the communication into a single dimension in a way that controls for the dependence in observations for a given sender and the amount of green cards drawn. We first code an honest representation as 0, overreporting as 1, and underreporting as -1. We then average these numbers over rounds to generate an individual-specific communication score. The higher the score, the more overreporting by the individual.

Figure 3 shows the resulting distribution of scores. We see a spike at 0, which includes honest subjects as well as a few subjects whose over- and underreporting exactly offset each other. The remaining subjects show a clear shift towards more negative scores (underreporting) in the *Public* treatment. The two distributions differ significantly on a Wilcoxon rank-sum test ($p = 0.0046$, two-sided). Thus, senders in the *Public* treatment do indeed act as if they are in need of justification.

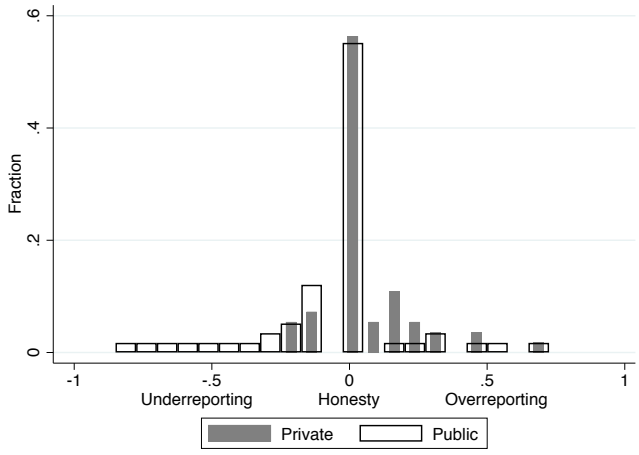


Figure 3: **One dimensional communication strategies.** An individual observation is the average communication of an individual over the interaction rounds, where honesty is coded as 0, overreporting as 1, and underreporting as -1.

Summary 1. *We find evidence for both under- and overreporting in both treatments. In line with Hypothesis 1, senders show about 10 percent fewer green cards in the* Public *treatment. This is due to both a drop in overreporting and a rise in underreporting, with the former being more pronounced.*

4.1 The justification motive

We now look deeper into the reasons behind the treatment difference in communication strategies. As explained in Section 3, the logic of the justification motive is that in the *Public* treatment, reporting a green card has a "price". The fact that actions are visible forces the sender to either incur the cost of a donation (choosing *Option 1*), or to reveal that she is not motivated to donate despite a high return of the donation (choosing

Option 2), appearing "hypocritical" in the terminology of FvW18. Hypothesis 2 thus specifies that more senders will follow up with a donation in this treatment.

Figure 4 shows the rates of prosocial behavior (choosing *Option 1*) among senders in both treatments, after reporting either a red or a green card. The fraction of prosocial decisions for each individual constitutes one observation. It is clear that in either treatment, few senders donate after showing a red signal. Senders are much more likely to donate in both treatments when they show a green card. Moreover, in line with Hypothesis 2, they do so more often in the *Public* treatment (64 percent on average) than in the *Private* treatment (42 percent), a difference which is significant on a Mann-Whitney ranksum test ($p = 0.0050$, two-sided). Thus, senders in the *Public* treatment do indeed act as if they face a higher cost of showing "frivolous" green cards.

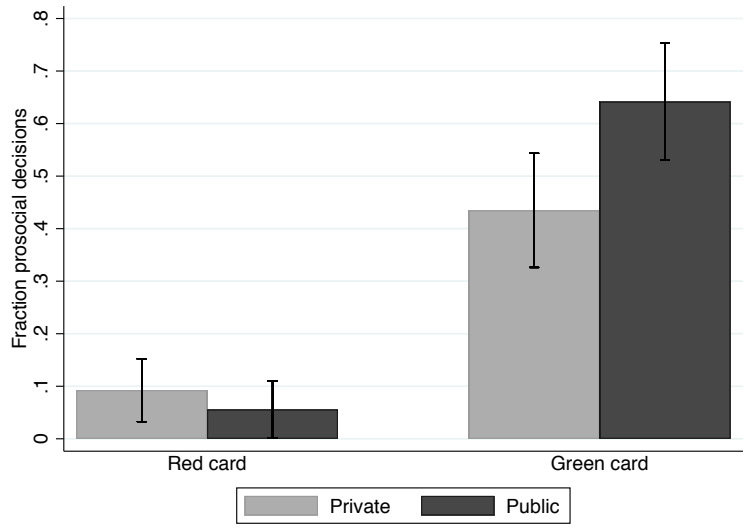


Figure 4: **Fraction of senders' prosocial decisions by card shown and treatment.** One observation is the fraction of prosocial decisions for each individual. Bars show 95 percent confidence interval, based on a normal approximation.

Relatedly, Hypothesis 3 predicts that senders who are less motivated to donate will be more likely to misreport a green signal as a red signal. While the analysis above already provides a correlation between prosocial behavior and the report of green cards, it is based on prosocial behavior that is measured post-treatment. To provide a stronger test of the theory, we correlate the fraction of underreporting with our pre-treatment measure of SVO.

Table 3 shows OLS regressions of the fraction of underreported green cards on the SVO score. Column (1) shows that across treatments, there is a clear negative correlation. Column (2) and column (3) split this result by treatment. The correlation is significant only in the *Public* treatment. Thus, our results support the interpretation that the justification motive plays a role mainly for those with low intrinsic motivation

who face strong image concerns.[14]

	(1) All	(2) Private	(3) Public
SVO angle	-0.00414**	-0.00129	-0.00507*
	(0.00167)	(0.00226)	(0.00222)
Constant	0.261***	0.144	0.316**
	(0.0688)	(0.0746)	(0.0996)
Observations	113	55	58

Table 3: **OLS regressions of fraction of underreporting on SVO score.** Column (1) shows the correlation for both treatments, column (2) for the *Private* treatment and column (3) for the *Public* treatment. Standard errors in parenthesis are clustered by session. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Summary 2. *In line with Hypothesis 2, senders in the* Public *treatment are much more likely to follow up a green card with a donation. In line with Hypothesis 3, underreporting is negatively correlated with SVO scores in the Public treatment.*

4.2 The persuasion motive

As noticed in the hypotheses section, overreporting does not benefit the sender monetarily. Even if it induces a donation by the receiver, there is only a one-third chance that the charity earns money, while there is a two-third possibility that the receiver loses €10. According to the theory by FvW18, overreporting thus requires a very high weight on the income of the charity and a very low weight on the receiver's income.

To see if such weights can explain the data, we look first at the fraction of senders who follow up on their overreport with a donation. More than half of senders who overreport never follow up with a donation, and no sender always follows up. In addition, in Table 4, we look at the correlation between the individual fraction of overreports and intrinsic motivation, as measured by the SVO score. We see no significant correlation overall, and a *negative* correlation in the *Public* treatment. Thus, it seems unlikely that overreporting is the result of an exceptionally high weight on the income of the charity.

A slightly different explanation is that making others contribute may generate some form of "warm glow", perhaps by reducing guilt for a sender who does not want to donate herself. This explanation differs from the previous one in that warm glow derives not from the effect of the donation, which is likely to be low, but from the act of inducing others to donate. Such feelings may be especially relevant if the sender thinks the charity is deserving of donations, and hence feels guilty about not donating herself.

[14]The motives behind underreporting in the *Private* treatment are somewhat unclear. This behavior cannot be explained by image concerns, as these are ruled out by design. We don't find clear patterns between a sender's underreporting, his/her SVO score and the rating of the "deservingness" of the charity (rated during the questionnaire by each subject on a 10 point scale). Rather, a part of the explanation may be that some subjects reported somewhat carelessly in the private condition. Three subjects write in the questionnaire that they reported "randomly", while one writes that the only round where (s)he underreported was a "mistake". Since no person engaged in this communication consistently, it thus seems unwise to seek too much behind it.

	(1) All	(2) Private	(3) Public
SVO angle	-0.00141	0.000116	-0.00316*
	(0.00152)	(0.00365)	(0.00141)
Constant	0.205***	0.230*	0.178**
	(0.0484)	(0.113)	(0.0573)
Observations	112	55	57

Table 4: **OLS regressions of fraction of overreporting on SVO score.** Column (1) shows the correlation for both treatments, column (2) for the *Private* treatment and column (3) for the *Public* treatment. Standard errors in parenthesis are clustered by session. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. There is one missing observation in the public treatment, as one sender never drew a red card, and hence did not have an opportunity to overreport.

To investigate this explanation, we look at the correlations between sender's average overreporting and his or her rating of the "deservingness" of the charity. We find a modest but significant correlation (Pearson $\rho = 0.20$, $p = 0.033$). By contrast, correlations of deservingness with individual measures of honesty and underreporting are negative and insignificant. In line with the idea that persuasion helps assuage guilt about a lack of personal contributions, the correlation is especially high among those who never follow up their overreport with a donation (Pearson $\rho = 0.52$, $p = 0.021$). Furthermore, five participants who always chose to show the green card state explicitly that their aim was to persuade the receiver to benefit the charity, even if they did not do so themselves. These results are suggestive of the idea that people derive psychological benefits from making others contribute to a charity they deem deserving, even if they don't contribute themselves. Clearly, more research is necessary to further test and confirm these conjectures.

Summary 3. *Overreporting does not correlate with a high willingness to donate. We find some suggestive evidence that inducing contributions by others generates psychological benefits.*

4.3 Impact of communication on receivers

Does communication matter for the behavior of receivers? To answer this question, we can simply compare the behavior of receivers who observe a green card with those who observe a red card. Figure 5 does just that, by showing the proportion of prosocial decisions for receivers, split by both treatment and by the card seen. It is obvious from the figure that communication matters: in line with Hypothesis 4, receivers are about 40 percentage points more likely to donate after being shown a green card, as measured within subject (Wilcoxon signed-rank test, $p < 0.001$). This tendency is slightly more pronounced in the *Public* treatment, which makes sense given that observing a green card is slightly more informative in this condition, but the difference-in-difference comparison is not significant (MWU ranksum test, $p = 0.28$).

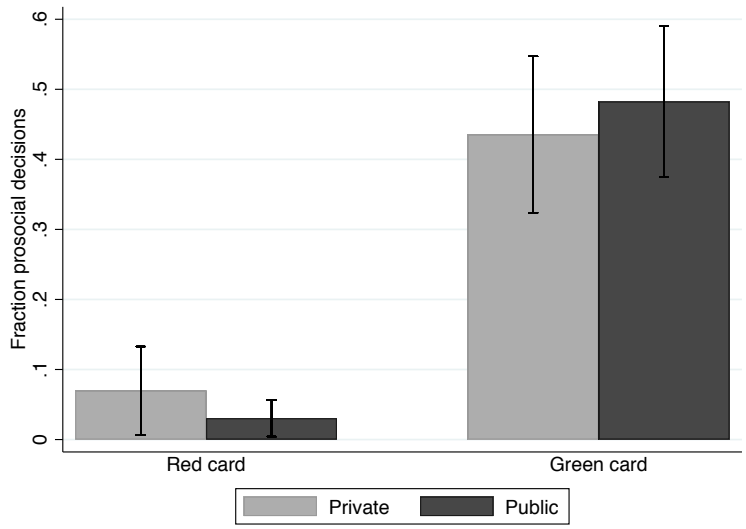


Figure 5: **Receivers' fraction of prosocial decisions by card seen and treatment.** One observation is the fraction of prosocial decisions for each individual. Bars show the 95 percent confidence interval, based on a normal approximation.

Impact of communication by state and SVO score. The true impact of the donation (i.e. the red or green state) determines the effects of communication on the receiver and the charity. While giving in the low impact state reduces the combined earnings of charity and donor, giving in the high impact state increases it. The left panel of Figure 6 shows the impact of the *Public* treatment on receiver giving. Clearly, this impact is negligible in the low impact state. In the high-impact state, there is a drop in the *Public* treatment. The size of the drop is about 4 percentage points or 11 percent, which is roughly in line with the 10 percent fall in the fraction of green cards shown demonstrated above, but the difference is not statistically significant.

To further understand the impact of communication in both states, we look at the SVO score of the receiver. Although this is a post-hoc conjecture, the impact of communication can be reasonably hypothesized to be higher for participants who are motivated to give to the charity, as indicated by their SVO score. Based on the classification in Murphy et al. (2011), we create a "low" and "high" SVO group of roughly equal size.[15] The middle and right panel of Figure 6 show the giving of high and low SVO individuals in both states.

Figure 6 shows several interesting patterns. First, as expected, high SVO individuals are more likely to donate in either state. Second, the impact of the treatment is very different for both groups. Contrary to expectations, low SVO individuals *increase* their giving in the *Public* treatment. This is somewhat surprising, but because samples are

[15]Murphy et al. (2011) propose thresholds for SVO values to classify individuals as "competitive", "individualistic", "prosocial" and "altruistic". To create a binary distinction, we merge the first two and the last two categories.

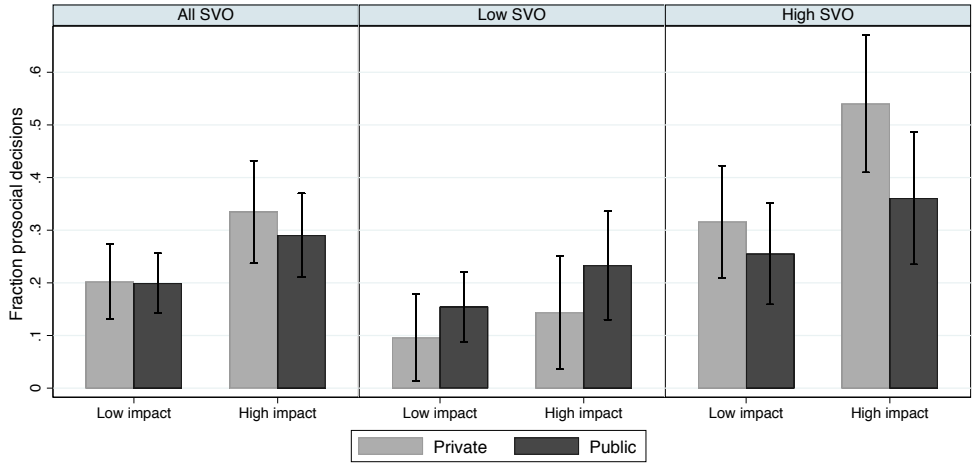


Figure 6: **Receivers' fraction of prosocial decisions by treatment and state of the world.** One observation is the fraction of prosocial decisions for each individual. Bars show the 95 percent confidence interval, based on a normal approximation. The left panel shows all data, the middle panel shows the patterns for "individualistic" and "competitive" individuals and the right panel for "prosocial" and "altruistic" individuals, using the SVO typology in Murphy et al. (2011).

small (approximately 28 individuals in each cell), these differences are not statistically significant. For high SVO individuals, patterns conform to the theory as giving shows a *decrease*, in both the low and high impact state. While the former drop is not statistically significant, the large drop in the high impact state (from 54 percent to 36 percent) is marginally statistically significant (Wilcoxon rank-sum, $p = 0.054$).[16]

Summary 4. *In line with Hypothesis 4, the color of the card shown by the sender has a big impact on giving rates of receivers. The Public treatment reduces giving by about 11 percent in the high impact state, with a sharper drop among givers with a high degree of intrinsic motivation (SVO).*

5 Discussion and conclusion

We investigate communication about the returns to prosocial actions. We find that although many subjects are honest, both under- and overreporting are prevalent, pointing at both persuasion and justification motives. Underlining the importance of justification, we show that increasing image or reputation concerns causes participants to communicate a high impact of a donation less often, in order not to look selfish.

These results show that communication about charitable giving is systematically distorted: in situations where actions are not observable and talk is cheap, participants "exaggerate" impact to persuade others to contribute. When talk is not cheap because

[16]These results are also borne out in regression analysis of receivers prosocial actions on a treatment and a state dummy and their interaction (adding random effects for receiver ID). We find a significant negative effect of the *Public* treatment on high SVO receivers in the high impact setting ($p = 0.012$).

donations are observable, participants "downplay" impact to justify their inaction. Both distortions occur despite the fact that senders derive no personal gain from miscommunication, and have a detrimental impact on the effectiveness of donations by the receivers of the communication.

Such misrepresentation can help explain the contentious nature of even factual observations in discussions of foreign aid, mentioned in the introduction. Moreover, the falsification of beliefs might play a role in a wide range of applications where reputation as a "good person" is at stake. For instance, in the literature on climate change, evidence from interviews and focus groups indicates that people deny the importance of climate change and their personal responsibility as consumers and voters, in order to avoid changes to their lifestyle (Stoll-Kleemann et al., 2001; Norgaard, 2006). Justification motives may thus help explain substantial minorities in many countries who do not believe in the scientific consensus on climate change.

While this experiment took place within the laboratory, there are several reasons to think that persuasion and justification motives have a stronger impact in other contexts. First, the signal in our experiment is relatively unambiguous. To the extent that signals outside the lab are multi-interpretable, people can more easily convince *themselves* of the truth of their falsifications. This may be particularly true for justification, in line with a well-documented tendency for ethical beliefs to be self-serving (Kunda, 1990; Gino et al., 2016). Second, the communication in our experiment was highly stylized, consisting only of binary signals. Natural languages offer much richer shades of persuasion and deception, potentially affecting participants who are not willing to lie outright in our experiment. Finally, image concerns in the lab are likely to be limited, as participants interact with strangers. Outside the lab, image concerns are of paramount importance in the political arena as well as on social networks, increasing the motive for downplaying social returns.

Future research can determine the relevance of justification and persuasion in different policy debates. There are also interesting extensions to be explored in the current design. For instance, what happens when senders face multiple receivers, increasing both the demand for justification and the consequences of (deceptive) reporting. Conversely, if receivers get input from multiple senders this may amplify the justification motive by giving senders shared, and perhaps diluted, responsibility for truthful reporting.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond, "Preferences for Truth-Telling," *Econometrica*, 2018, *In press.*

Andreoni, James and Alison Sanchez, "Do Beliefs Justify Actions or Do Actions Justify Beliefs? An Experiment on Stated Beliefs, Revealed Beliefs, and Social-Image Manipulation," *National Bureau of Economic Research Working Paper Series*, November 2014, *No. 20649*, 1–38.

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

_ **and Ragan Petrie**, "Public goods experiments without confidentiality: a glimpse into fund-raising," *Journal of Public Economics*, 2004, *88*, 1605–1623.

Angus Reid Institute, "What stops Canadians from donating more to charitable organizations," Technical Report, Angus Reid Institute 2017.

Ariely, Dan, Anat Bracha, and Stephan Meier, "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, mar 2009, *99* (1), 544–555.

Bénabou, Roland, Armin Falk, and Jean Tirole, "Narratives , Imperatives and Moral Reasoning," *NBER Working Paper 24798*, July 2018.

Bond, "UK Public Attitudes Towards Development: Aid Attitude Tracker Summary," Technical Report, Bond for international development 2015.

Butera, Luigi and Jeffrey Horn, "Good News, Bad News, and Social Image: The Market for Charitable Giving," *Working Paper*, November 2017.

Chatham House and YouGov, "British Attitudes towards the UK's International Priorities," Technical Report, Chatham House / YouGov 2011.

Crosetto, Paolo, Ori Weisel, and Fabian Winter, "A flexible z-Tree implementation of the Social Value Orientation Slider Measure," *Jena Economic Research Papers*, 2012, *62*, 1–8.

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

Eurobarometer, "Special Eurobarometer 222: Attitudes Towards Development Aid," Technical Report, European Commission 2005.

Exley, Christine L., "Excusing selfishness in charitable giving: The role of risk," *Review of Economic Studies*, 2016, *83* (2), 587–628.

_ , "Using Charity Performance Metrics as an Excuse Not to Give," *Mimeo, Harvard University*, 2016.

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

Foerster, Manuel and Joël J. van der Weele, "Denial and Alarmism in Collective Action Problems," *Tinbergen Institute Discussion paper*, 2018, *019*.

Gino, Francesca, Michael I. Norton, and Roberto A. Weber, "Motivated Bayesians: Feeling Moral While Acting Egoistically," *Journal of Economic Perspectives*, 2016, *30* (3), 189–212.

Gneezy, Uri, "Deception: The role of consequences," *The American Economic Review*, 2005, *95* (1), 384–394.

_ , Elizabeth A. Keenan, and Ayelet Gneezy, "Avoiding overhead aversion in charity," *Science*, 2014, *346* (6209), 632–5.

Gordon, Teresa P., Cathryn L. Knock, and Daniel G. Neely, "The role of rating agencies in the market for charitable contributions: An empirical test," *Journal of Accounting and Public Policy*, 2009, *28* (6), 469–484.

Grossman, Zachary and Joël J. van der Weele, "Self-Image and Willful Ignorance in Social Decisions," *Journal of the European Economic Association*, 2017, *15* (1), 173–217.

Harbaugh, William T., "What Do Donations Buy? A Model of Philanthropy Based on Prestige and Warm Glow," *Journal of Public Economics*, 1998, *67*, 269–284.

Karlan, Dean and Daniel H. Wood, "The effect of effectiveness: Donor response to aid effectiveness in a direct mail fundraising experiment," *Journal of Behavioral and Experimental Economics*, 2017, *66*, 1–8.

_ and Margaret A. McConnell, "Hey look at me: The effect of giving circles on giving," *Journal of Economic Behavior and Organization*, 2014, *106*, 402–412.

Kunda, Ziva, "The case for motivated reasoning," *Psychological Bulletin*, 1990, *108* (3), 480–498.

Kuran, Timur, "Preference Falsification, Policy Continuity and Collective Conservatism," *The Economic Journal*, 1987, *97* (387), 642–665.

_ , *Private truths, public lies: The social consequences of preference falsification*, Cambridge, MA: Harvard University Press, 1997.

Lacetera, Nicola and Mario Macis, "Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme," *Journal of Economic Behavior and Organization*, 2010, *76* (2), 225–237.

Metzger, Laura and Isabel Günther, "Making an impact? The relevance of information on aid effectiveness for charitable giving. A laboratory experiment," *ETH Zürich Working paper*, 2015, *08*.

Murphy, Ryan O., Kurt A. Ackermann, and Michel J.J. Handgraaf, "Measuring social value orientation," *Judgement and Decision Making*, 2011, *6* (8), 771–781.

Norgaard, Kari Marie, "People to Protect Themselves a Little Bit: Emotions, Denial, and Social Movement Nonparticipation," *Sociological Inquiry*, 2006, *76* (3), 372–396.

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

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

Stoll-Kleemann, S., Tim O'Riordan, and Carlo C. Jaeger, "The psychology of denial concerning climate mitigation measures: evidence from Swiss focus groups," *Global Environmental Change*, 2001, *11* (2), 107–117.

Yörük, Baris K., "Charity ratings," *Journal of Economics and Management Strategy*, 2016, *25* (1), 195–219.

Appendix

A Analysis of communication split by card drawn

Below we report the results for over- and underreporting in isolation. In summary, overreporting declines in the *Public* treatment, a result that is statistically significant at the 5 percent level in all comparisons. Underreporting increases in the *Public* treatment, and this difference is statistically significant at the 10 percent level in most, but not all statistical tests. Since all our p-values are either below 0.10 or 0.20, accepting the null hypothesis of no difference seems unwarranted, but more research is certainly desired to corroborate this result.

Non-parametric treatment comparison. We compute the fractions of over- and underreported cards by each individual. Comparing the fractions of overreporting, a two-sided Wilcoxon rank-sum test returns $p = 0.006$ and a two-sided t-test $p = 0.030$. For underreporting, a two-sided Wilcoxon rank-sum test returns $p = 0.16$ and a two-sided t-test $p = 0.077$. Note that using two-sided tests is quite conservative: given our explicit theoretical framework, we could have justified one-sided tests, thus halving our p-values.

Regression evidence. Table A.1 shows the result of random effect regressions with all available data points, where the dependent variable is whether a green card is shown. We report both linear regressions (columns 1 and 2) and probit models (columns 3 and 4), and use random effects to control for individual dependence. Columns (1) and (3) refer to underreporting, columns (2) and (4) to overreporting. It is clear that the drop in overreporting in the *Public* treatment is statistically significant, while the rise in underreporting is marginally significant.

	(1) Drawn Green	(2) Drawn Red	(3) Drawn Green	(4) Drawn Red
Public treatment (PT)	-0.0987*	-0.122**	-0.0885*	-0.136**
	(0.0513)	(0.0582)	(0.0513)	(0.0549)
Constant	0.896***	0.231***		
	(0.0369)	(0.0414)		
Observations	493	438	493	438

Table A.1: **Regressions of green card shown on treatment.** Columns (1) and (2) show a linear model with individual random effects, columns (3) and (4) show random effect probit marginal effect estimations. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

B Experimental materials

B.1 Instructions

Below we present the instructions that were used in the experiment. We have highlighted the parts that only appeared in the *Public* treatment as well as omitted page breaks.

INSTRUCTIONS

Welcome to this experiment. Please switch off your mobile phones and refrain from communication for the duration of the experiment. Please read the instructions carefully, they contain everything you need to know to participate. Whenever you have a question or a concern, please raise your hand and one of the experimenters will come to your desk and answer your question in private.

You will receive €6 for showing up to this experiment. You can earn additional money based on your decisions and the decisions of other participants. The experiment consists of two parts, each with multiple rounds of decisions. At the end of the experiment, the experimenter will randomly select one round from each of the two parts for payment, with each round being equally likely to be selected. We will pay you according to your earnings in the two selected rounds. The other rounds will not be paid. Thus, think carefully when making each decision, as it could be the one that will be paid. At the end of the experiment, your earnings in the selected rounds and your show-up payment will be paid to you anonymously in cash.

Depending on your choices, the experimenters may transfer a donation to a charity, GiveDirectly. GiveDirectly transfers money to very poor families in developing countries. Here is an excerpt from the website "GiveDirectly.org":

> "We use mobile payments technology to send your donations to extremely poor families in the developing world in the most capital efficient way currently possible. $0.91 of your dollar ends up in the hands of the poor. Our model is setting the benchmark for philanthropic efficiency around the world. We strive to promote a new approach to philanthropy that uses constant experimentation and analytical rigor to understand the most impactful ways to achieve positive outcomes."

During the experiment, we will show you identities and testimonials of people who have passed the screening of GiveDirectly, and are potential recipients of the donations in this experiment. Their photos and testimonials are taken verbatim from the website "GiveDirectly.org" (including typos).

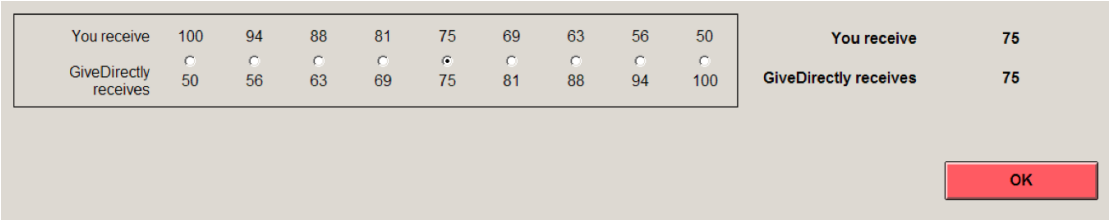
The experimenter will transfer the donation to GiveDirectly after the experiment. Note that the rules of the CREED laboratory do not permit deception of participants, so all promised donations will actually be made. If you want more information about the transfer, please contact the experimenter after this experimental session.

The instructions for the first part follow below. The instructions for the second part will be distributed after all participants have completed the first part.

[page break]

Instructions for Part 1

In this task you will be making a series of decisions on the computer. The decisions are about allocating resources between yourself and GiveDirectly. The screenshot below shows an example of such a decision, with earnings denoted in eurocents. Please make your choice by marking your most preferred allocation. In the example below, a person has chosen to distribute money so that (s)he receives €0.75 and GiveDirectly receives €0.75.



You will make six consecutive decisions that are similar to the example, but differ in the exact earnings. Your earnings in this part consist of the amount (in eurocents) that you allocated to yourself in the decision that is randomly selected for payment.

This completes the instructions for Part 1. When you have finished reading these instructions and if there are no questions, please click "NEXT" on your computer screen.

[page break]

Instructions for Part 2

In this part, participants are divided in two equally big groups of X participants and Y participants. Participants are randomly assigned to one of the types and will be identified with an ID: X1, X2, X3, ... etc. for X participants, and Y1, Y2, Y3, ... etc. for Y participants. We will communicate your type and ID on your screen at the beginning of this part.

In each round you will be seated across the table from a participant from the other group and engage in a short interaction. During the interaction we allow no communication between the two participants, other than the actions described in the task below. Verbal communication will result in exclusion from payment in this experiment.

At the beginning of the first round, both the X participants and the Y participants receive a decision sheet to record the decisions during the interaction. The decision sheet for each participant is private, and you should make sure it is not viewed by others.

Interaction type and communication

Each round, the experimenters roll a dice to determine the *type of interaction*. The type of the interaction can either be RED or GREEN: both are equally likely to be chosen. You do not learn which type of the interaction is chosen. However, before making a choice, Participant X will obtain some information about the interaction and can communicate with Y, as we now describe.

The experimenter will ask Participant X to take one card from a deck of three cards. When the true interaction is of type GREEN, the deck has two green cards and one red card. When

the true interaction is of type RED, the deck has two red cards and one green card. Thus, X's information is imprecise: with a chance of two thirds, X will see a card with a color that corresponds to the true type of the interaction in that round, while with a chance of one third, X will see a card with the wrong color. X writes down the color of the card on his/her decision sheet.

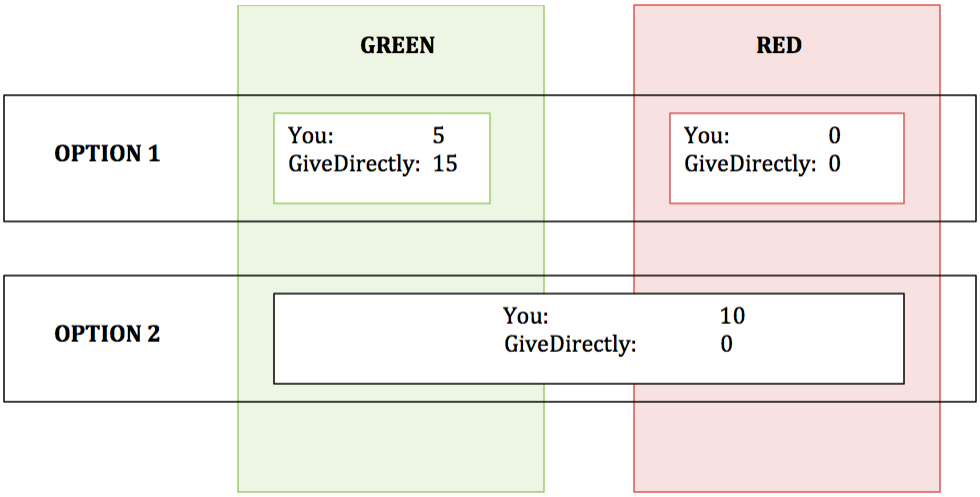
After Participant X has seen the card, X can communicate with participant Y, who has no information about the type of the interaction. To do so, X will find on his/her desk a RED and GREEN card, one of which she has to show to Y. Participant X is free to show any color to Y: it can correspond to the color of the card X drew earlier, or it can be a different color.

Choice

After participant X has shown the card, X and Y participants choose between Option 1 and Option 2, with the following payoff consequences.

- If you choose Option 1, earnings depend on the type of the interaction. If the type of interaction is GREEN, you will earn €5, and the experimenters will transfer a donation of €15 to GiveDirectly after the experiment is over. If the interaction type is RED, neither you nor GiveDirectly will receive anything from this choice.
- If you choose Option 2, you will receive €10, regardless of the interaction type (RED or GREEN). However, GiveDirectly will not receive any money in this case, even if the type is GREEN.

This table summarizes the payoffs.



The timing of the choices is as follows. First, both participants note down their choice on their private decision sheet. These choices are final and cannot be changed. Entries that are crossed out or corrected will not be paid.

[**(*Public* treatment only)** After both participants made their choice in private, Participant X reveals his or her choice by showing a card to participant Y with Option 1 or Option 2. Participant Y notes down this choice on her decision sheet. Participants will only receive payment

if X's choices on both decision sheets match: It is therefore important that Participant X reveals the choice honestly, and Y records it accurately.

Only Participant X reveals his/her choice, Participant Y's choice remains private.]

[page break]

Summary

Events in each round are as follows:

1. Each participant notes the ID of the other participant on his/her private decision sheet.
2. The experimenter randomly determines the type of interaction, RED or GREEN, where each has equal probability.
3. Participant X draws a card showing the correct color with two third chance, and writes down the color on the decision sheet.
4. Participant X chooses a RED or a GREEN card and shows it to Participant Y.
5. Both participants note down their choice between Option 1 or Option 2 on their private decision sheet.
6. [**(*Public* treatment only)** Participant X reveals his/her choice by showing the corresponding card of Option 1 or Option 2. Participant Y accurately records this choice.]
7. All X participants stay seated, and all Y participants move one space to their right, taking their decision sheets with them.

Please wait for a sign of the experimenter before you take each action. We remind you that no communication is allowed except for the actions just described.

[**(*Public* treatment only)** Public announcements of decisions and] payment

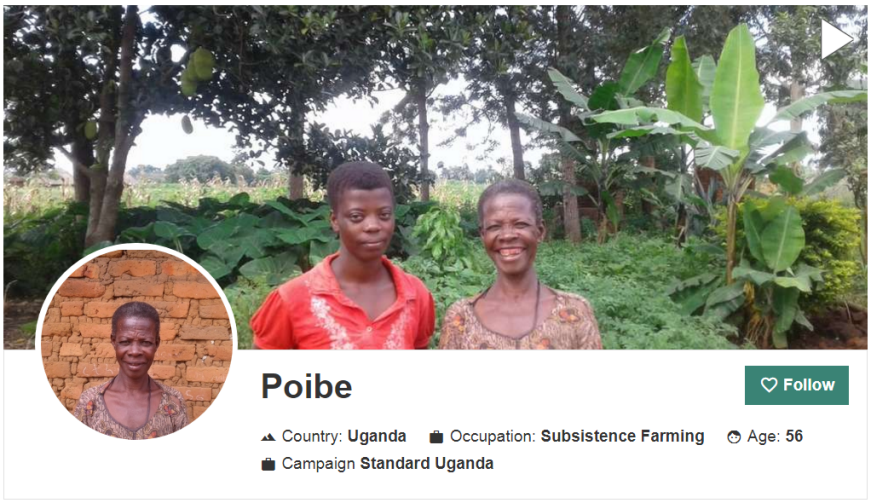
After the final round has concluded, the experimenter will select the round that is relevant for payment. The results of the other, not selected rounds will not result in any earnings.

[**(*Public* treatment only)** The experimenter will then ask all X participants to stand up one by one, and report in public the decisions (s)he has made in the selected round. Each X participant first reports the choice to show either a GREEN or a RED card to participant Y, and then the choice between Option 1 and Option 2. The X participant does *not* report whether (s)he saw a RED or GREEN card. Thus, *all participants* learn the decisions each participant X made in the relevant round.]

The experimenter will then collect the decision sheets, and ask you to go back to your original cubicle. While you answer a few questions on the computer screen, the payments are being prepared.

This completes the instructions for Part 2. If there are no questions, please click "NEXT" on your computer screen. When all are done reading, we will ask a few control questions to test your understanding of the procedures and the earnings. Afterwards, there will be a practice round to familiarize yourself with the procedures.

B.2 Reminder Sheet



What does receiving this money mean to you?

Receiving this money means i will use for paying school fees for my daughter in Aligoi secondary school. My daughter will study and get a better apaying ajob in future and it shall be of great help to the family.

Reminder of the payoffs

	GREEN	RED
OPTION 1	You: 5 GiveDirectly: 15	You: 0 GiveDirectly: 0
OPTION 2	You: 10 GiveDirectly: 0	