NORMS MAKE PREFERENCES SOCIAL

Erik O. Kimbrough
Simon Fraser University

Alexander Vostroknutov
Maastricht University

Abstract

We explore the idea that prosocial behavior in experimental games is driven by social norms imported into the laboratory. Under this view, differences in behavior across subjects is driven by heterogeneity in sensitivity to social norms. We introduce an incentivized method of eliciting individual norm-sensitivity, and we show how it relates to play in public goods, trust, dictator and ultimatum games. We show how our observations can be rationalized in a stylized model of norm-dependent preferences under reasonable assumptions about the nature of social norms. Then we directly elicit norms in these games to test the robustness of our interpretation. (JEL: C91, C92, D03)

Man is as much a rule-following animal as he is a purpose-seeking one.

(Friedrich Hayek 1973, *Law Legislation and Liberty, Vol 1: Rules and Order*, p. 11)

Without this sacred regard to general rules, there is no man whose conduct can be much depended upon. It is this which constitutes the most essential difference between a man of principle and honor and a worthless fellow. [...] (T)hat *reverence for the rule* which past experience has impressed upon him, checks the impetuosity of his passion, and helps him to correct the too partial views which self-love might otherwise suggest, of what was proper to be done in his situation.

(Adam Smith 1759, *The Theory of Moral Sentiments*, §3.5.2, *italics* added)

1. Introduction

Over the last thirty years, economists have uncovered robust evidence of human sociality in simple, anonymous laboratory games. Subjects display systematic tendencies towards egalitarian outcomes, cooperative strategies and reciprocal behavior, often in violation of the predictions of selfish profit maximization.

How to interpret these observations remains a lingering puzzle. Since the 1990s, there has been great progress in understanding sociality by modifying the utility function of individual players to incorporate explicit preferences over payoff distributions (i.e. social preferences), in the form of pure altruism and spite, inequity aversion, concerns about social welfare, and reciprocity.[1] In support of models defining

---

1.   See e.g. Rabin (1993); Levine (1998); Fehr and Schmidt (1999); Fehr and Gächter (2000); Bolton and Ockenfels (2000); Charness and Rabin (2002); Dufwenberg and Kirchsteiger (2004); Falk and Fischbacher (2006); Cox et al. (2007, 2008).

preferences over payoff distributions, there is evidence that many people trade off own and other's payoffs so that other-regarding behavior is sensitive to its 'price' (see e.g. Andreoni and Miller 2002; Fisman et al. 2007).

Despite the marked progress, there remain a number of empirical observations for which these models cannot account. First, there is evidence that minute changes to decision context can radically alter the nature and degree of prosociality observed in the lab. For example, dictator game giving is sensitive to knowledge about who is watching, to the process by which the right to be dictator is allocated, and to subtle manipulations of the choice set (Hoffman et al. 1994, 1996; Cherry et al. 2002; List 2007).[2] Second, recent evidence suggests that given the opportunity to conceal their choices, individuals who would typically choose to be generous, e.g. when forced into playing a dictator game, will instead choose selfishly—exploiting "moral wiggle room" (Dana et al. 2007; Andreoni and Bernheim 2009; Lazear et al. 2012).

In light of these findings, two new approaches have been adopted in the literature. The first considers a class of models in which prosocial decisions result from self- and other-signalling (Bénabou and Tirole 2006; Andreoni and Bernheim 2009). In these models, individuals care about their reputations and believe that others make inferences about their type on the basis of observed choices in experimental games. Intuitively, such models account for prosociality under the assumption that people's

---

2. See also Hoffman and Spitzer (1985); Andreoni (1995); Eckel and Grossman (1996); Burnham et al. (2000); Goeree and Holt (2001); McCabe et al. (2003); Bardsley (2008); Fershtman et al. (2012); Smith and Wilson (Forthcoming).

reputations are tied to their degree of prosociality, and they account for many of the observations noted above because the treatment interventions described disrupt the ability for observers to draw inferences about a player's type.

A second, complementary approach argues that prosocial behavior is driven by a desire to adhere to social norms (López-Pérez 2008; Kessler and Leider 2012; Krupka and Weber 2009, 2013; Cappelen et al. 2007).[3] Our paper follows this approach. The idea is that sociality is driven not directly by preferences over payoff distributions, but rather by preferences for following known social rules (be they written rules or informal norms). We refer to these as "*norms*" and assume that they specify the most socially appropriate action for a decision-maker in a given set of circumstances.[4,5] When people judge behavior, they compare it to an external, socially defined normative standard, and individuals internalize this process, judging

---

3.   This idea is well-known in psychology and sociology where early examples include Sherif (1936) and Merton (1957). Adam Smith's *Theory of Moral Sentiments*, captures this idea nicely with the notion that people evaluate their behavior in the view of an 'impartial spectator'.

4.   We have in mind "injunctive norms" which describe what ought to be done rather than "descriptive norms" which describe what people actually do (Cialdini and Trost 2008).

5.   Charness and Schram (2013) make a distinction between *social* and *moral* norms. They define social norms as those for which individuals seek "approval from one's peers" and thus involve common consensus on what is socially appropriate, and they claim that moral norms require introspection and are followed in order to avoid internal emotional reaction. We are skeptical about this distinction, but our paper emphasizes what they call social norms.

their own behavior according to its conformity to the norm.[6] This can be modeled

with a simple utility function in which deviations from norms generate a utility cost.

Then, when the norm is prosocial, people who suffer more from violating norms will

behave more prosocially. Under this line of argument, because prosocial norms exist

in many settings studied by experimenters, behavior that ultimately results from a

desire to follow social norms finds a proximate explanation in social preferences. It is

in this sense that 'norms make preferences social' - though they needn't always.[7]

Under this view, heterogeneity in sociality *across contexts* is due to the fact that

norms vary with context, as in the dictator games reported in List (2007), in which

measured aversion to inequity all but disappears. Thus, it is not an error or a violation

of rationality if people exhibit prosocial behavior in some contexts and not in others;

it is a natural consequence of the fact that people care about norms and that norms

are fundamentally context-dependent. A primary concern with this idea is that we

lack a theoretical foundation for identifying social norms in each context. A clever

---

6.   It may be useful to distinguish here between what we call a "norm" and what others have called

a "convention". Lewis (2002) introduced convention to game theory as an equilibrium solution to

a coordination problem (e.g. driving on the right side of the road).

7.   These approaches are complementary as a matter of epistemology; i.e. because norms are

implicit in any model of social preferences, reciprocity, or reputation. A model of fairness assumes *a

priori* agreement about what distribution constitutes (un)fairness. Similarly, as Charness and Rabin

(2002) put it, "any reciprocal model must embed assumptions about distributional preferences" (p.

824), and we cannot know which actions will help or harm a reputation without implicit reference

to norms. In a second sense, this is how norms make preferences social.

empirical take on this comes from Krupka and Weber (2013) who employ incentivized coordination games to directly elicit social norms in List's variants of the dictator game and then show that elicited norms can account for observed behavior in these games that is inconsistent with most models of social preferences. Their results suggest that if we can measure norms, even without a coherent model of how norms vary across situations, we can improve our understanding of social behavior.

A second implication of the social norms approach is that, when agents agree about the norm, observed heterogeneity in sociality *within a context* is a product of the fact that individuals differ in the degree to which they suffer from violating norms (Cappelen et al. 2007). To clarify this intuition, we formalize the idea that people suffer a disutility from violating norms, and importantly, that people differ in their sensitivity to own norm violations. We use an extremely simple framework to illustrate how norm-sensitivity may influence behavior in social settings, and then we report experiments designed to test for this relationship both within and between subjects.

To reveal the relationship between norm-sensitivity and prosociality, we develop a novel individual decision task (called the Rule-Following, or RF, Task) that measures subjects' preferences for following rules and norms, in a context that has *nothing* to do with social interaction or distributional concerns. Specifically, we tell subjects to follow a rule when doing so provides no monetary benefits and instead imposes monetary costs proportional to the time spent following the rule. Under these circumstances, only individuals who are intrinsically motivated to adhere to rules

and norms will follow the rule. Standard models of distributional preferences imply that everyone will break the rule because behavior only affects own payment. Unlike Krupka and Weber (2013) who measure only the average norm sensitivity in the population, our RF task provides a continuous measure of each *individual's* rule-following proclivity, and unlike Cappelen et al. (2007), our elicitation comes from an independent task, so the data are unconfounded with contextual cues present in many games that might influence measured sensitivity. We find evidence of extensive heterogeneity in the implied disutility of violating social rules.

We compare individual behavior in the RF Task to behavior in some of the most important games of sociality in the literature: public goods games, trust games, dictator games and ultimatum games. Norm-dependent utility has a central testable implication: a preference for following norms carries over from context to context, even if these contexts are unrelated. Thus, those who suffer more disutility from violating rules and norms (i.e. wait longer in the RF task) will be more likely to behave prosocially wherever there is a norm of prosocial behavior.

In our main treatment, we sort subjects, without their knowledge, into groups on the strength of their rule-following preferences, as measured by the RF task, and have them play standard repeated VCM public goods games. We find that assortatively matched high rule-following groups sustain cooperation with no evidence of decay, while low rule-following groups exhibit swift cooperative decay. We argue that our observations reflect a norm of conditional cooperation: rule-followers are willing to cooperate as long their group members do the same. Since this norm has less influence

on rule-breakers, those groups exhibit cooperative decay. In diagnostic treatments, we find no correlation between an individual's contributions and RF behavior when groups are *not* assortatively matched, thereby ruling out that the RF task merely measures other-regarding preference and validating the norm-based account of our findings.

In further treatments we use RF Task measurements to understand behavior in three other classic social preference experiments, allowing us to assess the extent to which our measure of norm-sensitivity is portable across games. We find that percentages returned (but not amounts sent) are significantly higher among assortatively matched groups of rule-followers in trust games. We also find that giving in dictator games and rejection thresholds (but not offers) in ultimatum games are higher among rule-followers. Importantly, behavior from our pooled sample for each experiment, ignoring RF heterogeneity, is comparable to that observed in the literature. Moreover, all of our findings can be shown to be natural implications of a model with norm-sensitive preferences under reasonable assumptions about the nature of the social norm in each setting.

Nevertheless, as noted above, we lack theoretical foundations that allow us to identify social norms *ex ante*. Thus, after conducting our main experiments and analysis, we performed additional experiments using the technique due to Krupka and Weber (2013) to directly measure norms of behavior in our games. Elicited norms are largely consistent with our interpretation, providing further support. Combining data on elicited norms and individual propensities to follow them (from

the RF task) we also report conditional logit analysis of subjects' choices under our utility specification. The analysis reveals that subjects with stronger rule-following preferences tend to put more weight on the norm component of their utility, further validating the framework of norm-dependent utility. Taken together, our experiments improve our understanding of the sources of prosocial behavior by showing how heterogeneous preferences for following norms can account for heterogeneous play in a variety of games.

Two questions may arise after reading the paper: "Where do norms come from?" and "How do individuals recognize the norm in a given setting?" One goal of this paper is to focus the discussion about prosociality on the importance of norms rather than individual preferences. If the reader accepts our argument that preferences are "made social" by reference to norms, then future research should seek to explain the creation, adoption and evolution of norms. A second important strand of research would aim to explain the origins of norm-sensitive preferences. A recent example providing an evolutionary foundation for one kind of norm-following can be found in Alger and Weibull (2013). We return to these issues in the conclusion.

2. Norm-Dependent Utility

In this Section we outline a simple norm-dependent utility specification that reflects the intuition behind our interpretation of the observed relationship between behavior in the RF task and in public goods, trust, dictator and ultimatum games – namely that heterogeneity in prosocial behavior is driven by differences in concern for social

norms that subjects import into the laboratory. Depending on the imported social norm in each setting, the implications of our utility specification for behavior will vary. We model a *norm* as a strategy profile: a norm describes the most socially appropriate choice for each player in each information set. A norm defines the "right" choice in each possible contingency, independent of any possible future decisions made by others.[8] For example, consider a norm of equity in the ultimatum game, which prescribes accepting relatively equal divisions of the pie and rejecting all sufficiently unequal offers. Such a norm provides a full description of socially appropriate actions at all potential choice nodes of the responder. For the proposer a norm might prescribe dividing the pie equally. In general, we assume both that all agents perceive a norm in each setting and that there is agreement about the content of the norm, but in practice, both of these assumptions may be satisfied to varying degrees. For the purposes of illustration, we prefer not to complicate matters, but we recognize that in practical terms both normative disagreement and weak norms may reduce the influence of norms on behavior. We return to this when we elicit norms directly.

Next we describe the norm-dependent utility function. Our work here builds on the model presented in Kessler and Leider (2012), though there are other ways to model norm-dependence that generate similar conclusions (see above). Briefly, norm-dependence implies a utility function which is increasing in own payoff and decreasing in the deviation between own action and the normative action. We introduce a norm

---

8. This aligns with Elster (1989) in which norms are "unconditional or, if conditional, not future-oriented," and López-Pérez (2008).

sensitivity parameter φ that models heterogeneity in norm following proclivity across individuals. Then, depending on assumptions about the normative action, it is easy to see how the model predicts differences in behavior. In an individual choice setting, those who are more sensitive to deviations from the norm will take actions more consistent with the norm. In games with a strategic component, choices may depend not only on own norm-sensitivity, but also on beliefs about the norm-sensitivity of others.

To illustrate norm-dependence in an individual choice setting, consider the standard Dictator game. Assume that the norm for the proposer is to give half of a pie (of size 2) to the Receiver. Then the proposer's norm-dependent utility can be specified as $U_p(x) = x - \varphi_p g(|x - 1|)$. Here, $x \in [0, 2]$ is the amount that the proposer keeps for herself, $\varphi_p \in [0, \infty)$ is the norm sensitivity and g is a strictly convex, differentiable, increasing function that represents the cost of violating the norm (deviating from the equal split).[9] The optimal choice is $x^*(\varphi_p) \in [1, 2]$ which weakly decreases in φ_p (see details in Appendix A.1): norm-followers with very high φ_p will give exactly half; norm-breakers with very low values of φ_p will give nothing; intermediate values of φ_p generate interior solutions. If instead the norm were to keep

---

9.   For illustration, we assume a common g and that people only differ in norm sensitivity φ. This allows us to connect behavior in the RF task (measure of φ) to behavior in social dilemmas. If g differs across individuals then we cannot separately identify g and φ (though see Cappelen et al. (2007) for an example of such identification). However, after conducting our main experiments we also elicited the average g for the population, and we report analysis treating this elicited g as the common one (see Section 4.6 below, which follows Krupka and Weber 2013).

everything (as might be the case when the proposer has *earned* the right to allocate the pie; see e.g. Bicchieri 2006), then high-φ and low-φ proposers would all keep the whole pie.

In games where two or more players move in sequence, we give a similar definition of utility. Suppose that in the Ultimatum game the norm for the proposer is to give half of the pie (of size 2) and for the responder to reject any offer less than half and accept otherwise. This completely specifies the norm (strategies for both players). We define the proposer's utility as $U_p(x, A) = x - \varphi_p g(|x - 1|)$ and $U_p(x, R) = -\varphi_p g(|x - 1|)$, where (x, A) represents the end node following the proposer's choice x and Acceptance, and (x, R) is the end node after x and Rejection. In either case, the proposer suffers disutility from violating the norm of equal sharing. The responder's utility from following the norm is $U_r(x, A) = 2 - x$ if $x \leq 1$ and $U_r(x, R) = 0$ if $x > 1$. If the responder violates the norm, then the term $-\varphi_r$ is added to both utilities $(U_r(x, R) = -\varphi_r$ if $x \leq 1$ and $U_r(x, A) = 2 - x - \varphi_r$ if $x > 1).$[10] In Appendix A.1 we show that in the Subgame Perfect Nash Equilibrium (SPNE) of the Ultimatum game with norm-dependent utility, the rejection threshold for responders weakly increases in φ_r. However, nothing certain can be said about a proposer's behavior: her offer depends on φ_p and her belief about the responder's rejection threshold. For example, if the proposer has low φ_p but believes that the responder is a higher type (and will only accept high offers), it is optimal for the proposer to offer the lowest amount that

---

10.   The norm-dependent utilities specified in this paper are special cases of a general specification for games with observable actions; see Kimbrough et al. (2014).

she believes the responder will accept. In this case the proposer's choice is completely determined by her belief about the responder's φ_r and independent of φ_p.

Similar arguments can be made for the other two games we study: trust and public goods games. As we review the data from our experiments, we will show how our observations can be rationalized as equilibria under norm-dependent utility given certain assumptions about the social norm in each setting (all details and proofs are in Appendices A.1, A.2, and A.4). Then, when we elicit norms directly, we find evidence that largely supports our interpretation.

3. Experimental Design and Hypotheses

The experiment consists of two decision-making stages and a questionnaire. Stage 1 of each treatment is the Rule Following task. This task allows us to measure a proxy for the parameter φ in the utility specification in Section 2. We then test whether higher-φ individuals are more likely to follow norms of cooperation than lower-φ individuals. Thus, in Stage 2 we compare the behavior of these types in a variety of games measuring social behavior. Our main treatment employs repeated VCM public goods games with fixed matching (PG). Then, we report repeated trust games with random rematching (TG) as well as one-shot dictator (DG) and ultimatum games (UG). Following Stage 2, participants complete a survey meant to measure their moral values. We describe each stage in detail below.

3.1. Stage 1 - The Rule Following Task

In Stage 1, the RF task, subjects control a stick figure walking across the computer screen. Each subject makes 5 decisions concerning how long they wait at a sequence of red traffic lights, each of which will turn green 5 seconds after their arrival. Figure 1 shows the screen that subjects see.

At the beginning of the RF task, the stick figure is standing at the left border of the screen, and all traffic lights are red.[11] Subjects initiate the RF task by pressing the START button. At this moment, the stick figure starts walking towards the first traffic light. Upon reaching the first red light, the stick figure automatically stops. The light turns green 5 seconds after the figure stops; however, subjects may press a button labeled 'WALK' any time after the figure stops. When a subject presses WALK, the figure continues walking to the next light before stopping again, and subjects must once again press WALK to continue to the next light. Throughout the RF task, the WALK button is shown in the middle of the screen. Subjects can press the WALK button at any time during the RF task. However, it is functional only when the figure stops at a traffic light.

Subjects receive an endowment of €8, and they are told that for each second they spend in the RF task they will lose €0.08. It takes 4 seconds to walk between each light, and 4 seconds from the final light to the finish. Therefore, all subjects lose

---

11.   Before subjects start the task, they see a short cartoon in which the traffic lights blink from red to green. This ensures that subjects understand that the lights can turn green.

FigURE 1. Screen shot of the Rule Following (RF) task.

around €2 walking, and if a subject waits for green at all 5 traffic lights, she will lose

an additional €2 waiting. Thus the most a subject can earn in the RF task is €6 (if

she spends no time waiting at lights), and the most she can earn if she waits is €4

(if she waits exactly 5 seconds at each light). In the instructions for the RF task (see

Appendix B.1) subjects are told: "The rule is to wait at each stop light until it turns

green." No other information, apart from the payment scheme and a description of

the walking procedure, is provided in the instructions.[12]

The rule following task creates a situation, familiar to most subjects, in which

they are asked to follow a rule at some cost to themselves. Waiting at a stoplight

when there are no other vehicles or individuals in sight is an example of seemingly

'irrational' obedience, in the sense that (barring traffic cameras) there is no cost to

breaking the rule. In such circumstances, the usual justification for obeying traffic

law—ensuring the safety of drivers and pedestrians—has no bite because there are no

other drivers or pedestrians to protect or be protected from. Yet in our experience,

---

12.   If subjects asked what would happen if they pass through the red light, an experimenter

explained that all information relevant to the experiment is in the instructions.

it is quite common for people to stop and wait impatiently at traffic lights, even in the middle of the night. We argue that norm-dependent preferences provide the explanation. Individuals who care about norms (or rules) will wait; their disutility from violating social expectations is greater than the utility from quickly getting to the destination, and others who are not so concerned (or who face large opportunity costs of waiting) will run the light.

In the lab, we control the opportunity cost of obedience, and by observing individual willingness to follow our costly rule, we measure a proxy for the parameter φ. Then we use this measure of φ to better understand observed decisions in four well-known experimental games in which behavior is plausibly guided by social norms.[13] Importantly, since all costs and benefits are private in the RF task, social preference

---

13.   Before making decisions in the RF task, subjects only receive instructions for that stage. They are aware that the experiment will consist of several stages, but they know neither what they will do in the next stage(s) nor the connection between the RF task and consecutive stages. In particular, subjects see a label that reads "Part 1" at the top of the rule following instructions (see Appendix B.1). In previous dictator game experiments, knowledge of the existence of an unspecified second-stage has been shown to alter subjects' behavior by making them more cooperative in expectation that their first-stage behavior may influence their second-stage reputation (Smith 2008). If subjects are concerned for their reputation and thus wait longer than they might in a treatment without an implicit 'shadow of the future' (or, similarly, with a double-blind protocol), this could only bias behavior in one direction. Any such 'strategic' rule-following would only yield false positives, diluting the information content of the RF task and strengthening any results we obtain that confirm our hypotheses.

models defined over payoff distributions have the same prediction as the selfish model
- all predict that everyone will break the rule.

3.2. Stage 2 - Games of Sociality

3.2.1. Public Goods Treatment.   In our PG treatment, subjects play 10 periods of a
repeated public goods game with a voluntary contributions mechanism in fixed groups
of 4 (Isaac and Walker 1988). Each subject receives an endowment of 50 tokens at the
beginning of each period (1 token = €0.01), and she must divide her tokens between
a group account and a private account. In each period, each subject earns the sum of
the amount placed in the private account plus the individual return from the group
account, which is 0.5 * (sum of all contributions). Thus, it is individually optimal
to contribute nothing to the group account and Pareto optimal for all subjects to
contribute their entire endowments. After each period, subjects learn their earnings
in that period, the sum of group account contributions from all members of their
group, and their total earnings through that period. Subjects are informed only that
they will participate in 'several' periods of decision-making.

Crucially, *unknown to the subjects*, their decisions in the RF task determine with
whom they are grouped in the PG stage. First, we randomly divide subjects into
groups of 8 (sessions consisted of 16, 24 or 32 subjects). Second, within each group of
8, we rank subjects according to the total time they spent waiting in the RF task – at
least 25 seconds for those subjects who waited for the green light at all traffic lights
and close to 0 seconds for those who did not wait at any light. Then, in each group

of 8, we separate the top 4 subjects (rule-followers) and the bottom 4 subjects (rule-breakers) into two groups for stage 2. After we match subjects, there is no interaction between any groups of 4. Subjects are not informed about the matching procedure, and they are told only that they will now interact with a fixed group of three other participants (see Appendix B.2).[14]

3.2.2. Trust Game Treatment.   In stage 2 of the TG treatment subjects are also sorted (without their knowledge) into groups of 4 on the basis of their RF task behavior, and then they play the trust game 6 times (Berg et al. 1995). Each subject plays the game twice with each other subject in the group. The order is randomized, and subjects receive no identifying information about their partner. Each subject participates 3 times in the role of first mover (blue person) and 3 times as a second mover (red person, see Appendix B.3). Subjects are informed only that they will make several decisions, but they are aware that they will participate in both roles.

Each subject receives an endowment of 80 tokens in each period (1 token = €0.01). The first mover sends any amount between 0 and 80 tokens, knowing that the amount sent will be multiplied by 3 and given to the second mover. The second mover then sends back to the first mover any amount between 0 and the amount received. In each period the earnings of the first mover are (80 tokens - tokens sent to the second

---

14.   Note that we did not deceive our subjects. None of the statements in the instructions are false or misleading. It is a separate, and also interesting, question whether subjects' behavior would change if they had knowledge of the sorting procedure, but our purpose was to use isolated rule-following behavior to identify subjects' types.

mover + tokens sent back from the second mover). The earnings of the second mover are (80 tokens + tokens received from the first mover - tokens sent back to the first mover). After each period subjects observe the amounts sent, received and returned in that period as well as their total earnings through that period.

3.2.3. Dictator and Ultimatum Game Treatments.   In stage 2 of the DG treatment, subjects play one round of the dictator game due to Forsythe et al. (1994), where one-half are randomly assigned to be dictators. In stage 2 of the UG treatment, subjects play one round of the ultimatum game due to Güth et al. (1982), where we elicit responder decisions via the strategy method.[15] Instructions are in appendices B.4 and B.5.

In these treatments, the RF task has no bearing on the second stage. Subjects are randomly assigned to be proposers who are allocating a pie worth €16. In the DG treatment, the proposer's offer (x) is final: the proposer receives $16 - x$, and the responder receives x. In the UG treatment, while the proposer is choosing how much to offer to the responder (y), the responder also chooses the minimum offer he would be willing to accept (y^*). If $y \geq y^*$, the proposer receives $16 - y$ and the responder receives y; otherwise, both receive nothing.

---

15.   In a meta-study, Brandts and Charness (2011) find that in experiments involving punishment or rejection, strategy method generally gives different results than standard extensive-form approach. Nevertheless, we are making comparisons between rule-followers and rule-breakers who all used strategy method, and we elicit norms for the strategy method version of the game. We do not have a reason to expect that extensive-form procedure would bias these two groups differentially.

3.2.4. Diagnostic Treatments. As robustness checks meant to rule out some alternative hypotheses, we also ran the following diagnostic treatments, which we will discuss in Sections 4.1 and 4.2:

1. A NoSort-PG treatment in which subjects first performed the RF task and then played the PG game with 3 randomly chosen individuals

2. A Reverse-PG treatment in which the PG game was played first with random matching, followed by the RF task and the questionnaire

3. A NoRule-PG treatment in which the phrase "The rule is to wait at each stop light until it turns green" in the instructions for the RF task was replaced by "5 seconds after the stick figure reaches a stop light, it will turn from red to green"

4. A NoRule-Reverse-PG treatment combining (2) and (3)

The NoSort and Reverse treatments, which use unsorted groups in the PG game, provide a baseline against which we compare the path of public goods contributions when groups *are* sorted according to RF behavior. As we discuss in Section 3.4 below, this allows us to better understand the norms underlying behavior by observing mixed-type groups. The NoRule treatments, on the other hand, allow us to determine what portion of observed rule-following is due to the the statement that "the rule is..." and what portion is due to the induced context.

3.3. Procedures

All subjects participated in both the RF task and *one* stage 2 game. Thus, we can make within-subject comparisons between the RF task and behavior in a single game.

We vary the stage 2 games between-subject to avoid concerns that experience in one strategic setting could influence play in another. After stages 1 and 2, subjects answered the Moral Foundations Questionnaire.[16] Then they received cash equal to the sum of earnings in stages 1 and 2. We substituted earnings from the RF task for a formal show-up payment. Table 1 summarizes our experimental design, including the number of subjects (and independent groups) participating in each treatment.

We also ran but do not report our first TG session, which fell prey to a software error, and two extensive form ultimatum game sessions (contra the reported sessions which use the strategy method). All data are available from the authors. No other data were collected for this experiment either in the form of pilots or other sessions/treatments. All experiments were programmed in z-Tree (Fischbacher 2007) and were conducted at Maastricht University's BEELab between May 2011 and February 2013.

3.4. Hypotheses

Our RF task allows us to classify subjects according to φ by observing the extent to which they incur costs in order to follow a rule. Norm-dependent utility implies that

---

16.   This is designed to measure the strength of subjects' respect for various moral values (Haidt and Joseph 2004; Graham et al. 2008, see Appendix C). While the list is not necessarily exhaustive, the purpose is to measure moral intuitions about five values: 1) aversion to doing *harm*; 2) concerns for justice or *fairness*; 3) love of country, family, and the *in-group*; 4) respect for *authority*; and 5) the desire for cleanliness and *purity*. Subjects answer 6 questions about each value using a Likert scale. We construct a score between 0 and 30 that represents the strength of respect for each value.

Treatments:	PG	NoSort	NoRule	Reverse	NoRuleReverse	TG	DG	UG
Stage 1	RF	RF	NoRule	PG	PG	RF	RF	RF
Stage 2	PG	PG	PG	RF	NoRule	TG	DG	UG
Post Experiment			Moral Foundations Questionnaire					
Group Size	4	4	4	4	4	4	2	2
Sorted (Y/N)	Y	N	Y	N	N	Y	N	N
# of Subjects	72	64	24	48	24	96	134	138
# of Groups	18	16	6	12	6	24	67	69
# of Obs. per Group	10	10	10	10	10	6	1	1

RF - Rule-Following Task, NoRule - RF task with no rule, DG - Dictator Game
UG - Ultimatum Game, TG - Trust Game, PG - Voluntary contributions PG

TABLE 1. Summary of Main Experimental Design

agents with high values of φ will be more inclined to behave in accordance with social norms of cooperation, generosity, and reciprocity than those with lower values of φ.

In our experiments, we do not allow subjects to discuss strategies, and we do not provide contextual cues in the instructions meant to induce particular norms. The success of using our RF task as a screening mechanism relies on the subsidiary hypothesis that subjects import norms of behavior from outside of the lab that influence their decision-making. Depending on the imported social norm in each setting, the specific hypotheses under our utility specification will vary. In Appendix A, we detail a set of possible outcomes in each game under reasonable assumptions about the relevant social norms.

Perhaps the least controversial example is the dictator game, where it seems reasonable based on observed behavior in prior experiments to assume that there is a norm of equal sharing. Given this norm, we would hypothesize that individuals with higher RF waiting times (high-φ) would give amounts closer to the equal split

than those with lower waiting times. This tradeoff between own payoff and following the norm rationalizes the typical distribution of DG giving.

In the PG game, it is easier to see how the hypothesized relationship between RF waiting times and behavior could vary. If the norm is to contribute *so long as others also contribute*, then only groups with members who all have sufficiently high φ will sustain cooperation (see Appendix A.2). Under an alternative norm that says giving *per se* is socially appropriate, independent of others' choices, we might also expect contributions to be higher among rule-followers - but this should be true independent of the matching scheme. Comparing the behavior of sorted groups of rule-followers and rule-breakers to behavior of unsorted groups in our NoSort and Reverse treatments will allow us to disentangle these competing explanations.

Similar arguments can be made for the TG and UG, in which, depending on assumptions about the norm, we would expect different relationships between RF behavior and play in the game. In the UG, if there exists a norm of sharing, as in the DG, then we might expect to see a similar difference in amount sent between high and low-φ proposers. However, as noted in Section 2, norms may exist for behavior of both proposers and responders. As second-movers, responders' decisions should be determined by their φ and the norm, and if proposers know this, then their "generosity" may be driven by strategic concerns about the norm-following behavior of the responder (i.e. her willingness to reject) rather than own concern for norms of generosity. If this is the case, then we would expect little difference in proposer behavior between types. A similar effect should exist for the proposers in the TG. If

they expect responders to follow the norm of reciprocity, by which we mean returning a high percentage of what was sent, then even low-φ proposers might decide to send non-zero amounts.

Our experimental design will allow us to observe play of each type and thereby to infer the norm from observed behavior. In our results section we discuss whether and how our observations can be rationalized under our utility specification. Then, to provide an additional test of our interpretation, we also conduct experiments that directly elicit subjects' beliefs about the social norm in each setting.

3.5. Norm Elicitations

We employ methods due to Krupka and Weber (2013) to directly elicit norms for the games described herein. Subjects were presented with a description of a scenario faced by players in one of the games we study, and they were instructed to rate the "social appropriateness" of various possible actions that the player could take. Their answers were incentivized using a coordination game where we randomly chose one of the hypothetical actions we asked them to rate and paid them €8 only if their evaluation was the same as the modal evaluation of peers in their experimental session. As discussed in Krupka and Weber (2013), this method nicely captures the notion of a norm as a set of shared beliefs about the appropriate course of action in a given scenario. If an action is rated "very socially appropriate" by all (or nearly all) people, then that action can be considered a norm. The weaker the agreement, the weaker the norm.

In each session we elicited norms of behavior for one player role in two games. Two sessions asked subjects to rate the social appropriateness of possible contributions in the one-shot Public Goods game and of possible amounts sent in the Dictator Game. Two sessions elicited norms at time t in the repeated Public Goods game, *conditional on others' contributions at time $t - 1$*. Two sessions elicited norms of sender behavior in the Trust and Ultimatum Games, and finally, two sessions elicited norms of receiver behavior in the Ultimatum and Trust Games. We took care to present each choice faced by the hypothetical chooser in the same manner as it was presented to subjects in our main experiments. Appendix D reproduces instructions and screenshots for each elicitation. In total we elicited norms from 205 subjects using z-Tree at Maastricht University in July and September 2014.

If the norms we elicit are consistent with the behavior we observe among rule-following types, this provides additional evidence in favor of our account of the findings. We return to this below in our discussion of the main results.

4. Experimental Findings

In this section we first analyze data from the RF task for all 600 individuals who participated in the main experiment. Then to test our hypotheses, we analyze our public goods treatment as well as diagnostic treatments meant to evaluate competing explanations. Finally, we report the results of our TG, DG, and UG treatments to test the robustness of our findings.

4.1. Behavior in the Rule-Following Task

The choice data important for our hypotheses is the number of seconds that each subject spends waiting at the traffic lights. This number is directly proportional to the money that the subject gains by not waiting (she can gain between €0 and €2). Suppose that the norm prescribes to wait at all lights (to gain €0). Then, the optimal waiting time under our utility specification can be computed from the maximization problem $x^* = \text{argmax}_{x \in [0,2]} x - \varphi g(x/2)$. Here x is divided by 2 so that the argument of g ranges from 0 to 1. The optimal x^* solves $g'(x^*/2) = 2/\varphi$. Thus, since g is strictly convex and increasing we can conclude that x^* (which we observe) and φ are monotonically related, which allows us to use x^* as a proxy for φ.[17]

Figure 2a displays a histogram of waiting times in the RF task. Average waiting time is 22.5 seconds. Notably, when the rule is invoked, 62.5% of subjects spend at least 25 seconds waiting, indicating that they obey the rule without exception, though it costs them at least €2.

---

17.   Another way to view obedience in our RF task is as a pure "experimenter demand" effect. Under that interpretation, we are using demand effect sensitivity as a proxy for φ. This has the nice feature that a long-time bogeyman of experimenters turns out to be an ally. We are sympathetic to this view, but we would argue that any experimenter demand effect is actually a manifestation of the norm-dependence we seek to measure, else why should individuals be concerned about the demands of the experimenter? Levitt and List (2007) suggest that demand effects may be responsible for much of the apparently anomalous heterogeneity in prosocial behaviors (e.g. context effects); in these cases too, we argue that sensitivity to a demand effect actually reveals norm-dependent preferences.
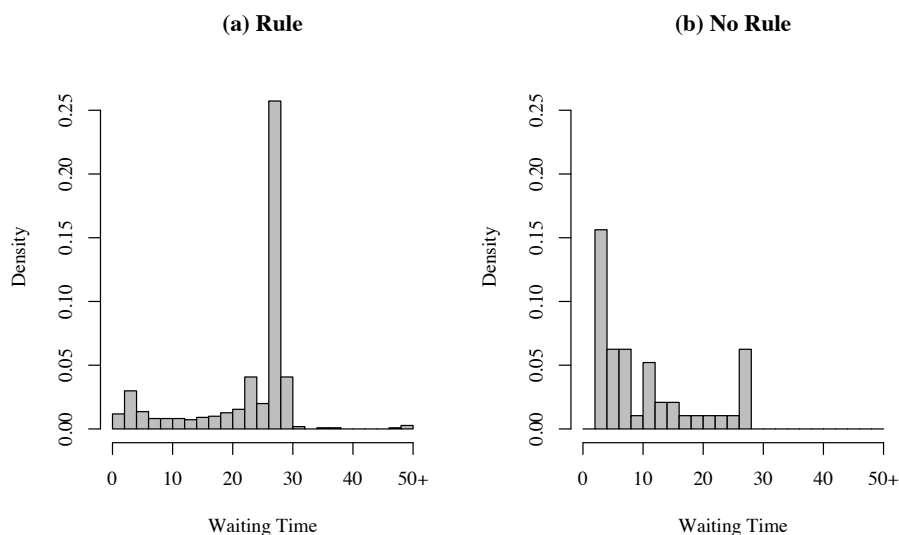
(a) Rule   **(b) No Rule**

FIGURE 2. Histograms of Waiting Time in Seconds, Rule vs. No-Rule Treatments

In the RF task, we induce a familiar traffic light context – we were concerned that otherwise our 'rule' would be ignored. However, because of the induced context it is not immediately clear how much observed obedience is induced by the statement that "the rule is. . ." and how much by imported norms associated with traffic lights. For this reason we conducted an additional "NoRule" treatment - described in Section 3.2. Figure 2b displays a histogram of waiting times in the NoRule treatment. Invoking a "rule" has a powerful impact on individual waiting times. In the NoRule treatments, average waiting time is only 10.4 seconds, and only 12.5% of subjects wait at least 25 seconds. This suggests that the induced context of our RF task is responsible for some of the rule-following behavior we observe, but explicit statement of the rule plays a more important role. Regression analysis in Appendix E.1 indicates that this

treatment difference is highly significant (p-value < 0.01).[18] Next we describe how we use RF data to better understand behavior in our stage 2 games.

In two treatments (PG and TG) we use RF task behavior to sort subjects into groups, and then we compare behavior across the two types of group. In each session, subjects were randomly assigned into groups of 8, and then those groups were split into 'rule-following' and 'rule-breaking' groups at the within-group median waiting time. Mean waiting time in rule-breaking groups is 19s (se $= 0.98$), while mean waiting time in rule-following groups is 27s (se $= 0.36$). Relatively high waiting times in rule-breaking groups are explained by the fact that many individuals classified as rule-breakers (because they were in the bottom four in their group of eight in the RF task) nevertheless followed the rule. Recall that only 37.5% of subjects broke the rule at all, so there is considerable noise in group assignment. Moreover, others classified as rule-breakers followed the rule at as many as 4 lights.[19] This suggests our classification is noisy and that observed effects thereof might be thought of as lower bounds. Nevertheless, our design allows us to test for the effects of assortative matching on RF behavior to identify whether groups of rule-followers are more likely to follow social norms in these games.

---

18. Figure E1 in Appendix E.1 also displays empirical CDFs of RF task behavior in each treatment.

19. The relatively high waiting times among rule-followers (> 27s) can be explained by reaction times - a lag of ~ 0.5s between the light changing and clicking "Walk."

In our other main treatments (DG and UG), we do not sort subjects but instead compare DG and UG behavior across various percentiles of the distribution of the RF task. Next we report data for each game in sequence.

4.2. Public Goods Treatment

Figure 3 displays time series of mean percent contributed by each group and associated standard errors in rule-following and rule-breaking groups. From the figure, it is clear that contributions decline over time only among rule-breaking groups; rule-following groups avoid the well-known pattern of cooperative decay. Despite the noise in group assignment such that 'rule-breaking' groups often contain at least one individual who followed the rule completely, on average, rule-following groups contribute 17% more of their endowment to the public good than rule-breaking groups over the entire experiment, and the difference is even larger (26%) in the second half of the experiment.

Regression analysis of mean group contributions in each period provides statistical support. Because we have panel data at the group level, we employ mixed-effects regression with random effects for each group, and we cluster standard errors at the group level. We regress group mean contributions on an intercept, a period trend, a rule-following group dummy and a period×rule-following dummy. Table 2 reports the results.

A negative and significant coefficient on the period trend indicates that contributions decline over time among rule-breakers, and a positive and highly
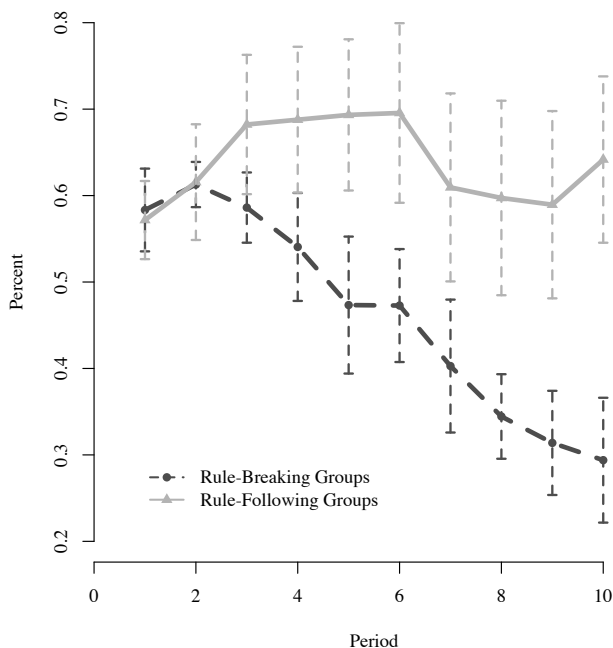
FIGURE 3. Time series of mean percent of endowment contributed ± 1 SE, for rule-following and rule-breaking groups in the PG treatment (computed at the group level; 9 independent observations underlying each line).

PG Contributions	Coef.	Std.Err.	z-value	Pr($> |z|$)
Intercept (Rule-Breakers)	33.651	1.345	25.02	0.000***
Period	-1.915	0.395	-4.85	0.000***
Rule-Following Group	-1.381	2.979	-0.46	0.643
Rule-Following Group $\times$ Period	1.851	0.613	3.02	0.003***

N = 180; 18 groups $\times$ 10 periods; standard errors clustered
at the group level. $* - (p < 0.10), ** - (p < 0.05), *** - (p < 0.01)$

TABLE 2. Mixed Effects Estimates of Group Mean PG Contributions (out of 50 tokens)

significant coefficient on the period$\times$rule-following interaction indicates that this effect is offset among rule-followers. A Wald test cannot reject the null hypothesis that these terms sum to 0 (p-value = 0.892), which supports the observation that

contributions do not decline among rule-followers.[20] In Appendix E.2.1 we provide non-parametric tests that support the same conclusions.

These data are consistent with the conjecture that the norm in the repeated PG game is one of *conditional cooperation*. Such a norm might require a player to contribute some amount η in period 1 and to continue contributing η each period *so long as others have done the same*. However, if others violate the norm, then it becomes appropriate to contribute nothing. In Appendix A.4 we show that under such a norm, contributions of η can be sustained as a Perfect Bayesian Nash Equilibrium if all players' own norm sensitivity parameters φ and *beliefs* about other players' norm sensitivities are sufficiently high. Such an equilibrium rationalizes our finding that groups consisting of rule-followers sustain cooperation while groups of rule-breakers do not.[21]

Finding 1: *Rule-following groups sustain significantly higher contributions than rule-breaking groups in the VCM public goods game.*

---

20. If we also include the 6 groups from our No-Rule treatment in the analysis, which were also sorted according to waiting time, the results remain essentially unchanged. See Figure E2 in Appendix E showing time series including sorted NoRule sessions in the computation of rule-following and rule-breaking means.

21. This model can also help explain why contributions in the first period are similar among rule-followers and rule-breakers. If rule-breakers believe that others have sufficiently high φ, then strategic incentives may encourage them to contribute early in order to induce contributions by others.

To better understand Finding 1, we report treatments designed to test its robustness and compare our norm-based explanation to other competing explanations. For example, one potential concern is that our rule-following task simply measures other-regarding preferences - people who follow rules are altruists. If this is true, then it is no surprise that rule-followers are more cooperative in Public Goods games. To distinguish these hypotheses, we can exploit our NoSort and Reverse treatments in which subjects were not sorted into groups.

If our rule-following task captures other-regarding preferences—instead of norm-dependence as we hypothesize—then the contributions of rule-breakers should systematically differ over time from those of rule-followers, even in the absence of sorting. We test this hypothesis using our data from 16 groups of 4 in the NoSort-PG treatment and 12 groups of 4 subjects in the Reverse-PG treatment (in which the order of stages was reversed so subjects were also unsorted).

Figure 4 shows time series of mean group-level contributions for the NoSort-PG and Reverse-PG treatments as well as for rule-following and rule-breaking groups in the PG treatment. When subjects are matched randomly into groups, the well-known pattern of cooperative decay reappears. In period 1 of the Reverse-PG and NoSort-PG treatments, the mean contribution is 60% of the endowment, and, in the PG treatment both rule-followers and rule-breakers average 58%. However, by period 10, Reverse-PG contributions decline to 41% of the endowment and NoSort-PG contributions fall to 43%, while rule-followers contribute 64% and rule-breakers

contribute 29%. Regression analysis in Table E3 in Appendix E supports the evidence presented in Figure 4.
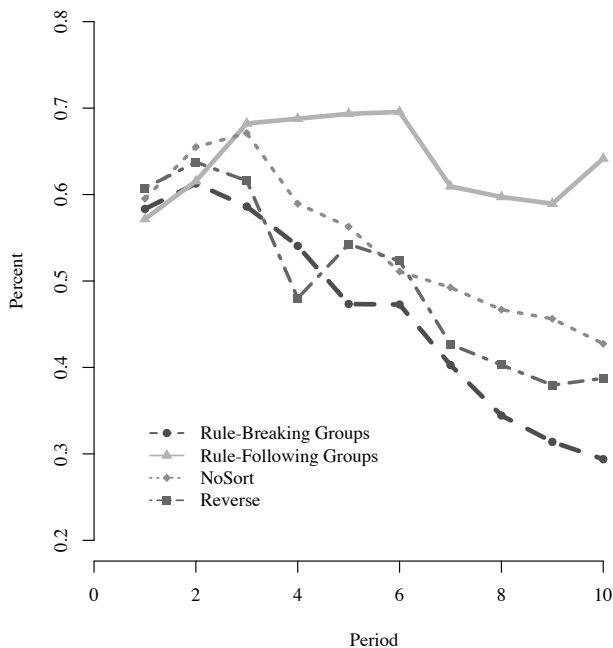

FIGURE 4. Time Series of Mean Group Public Good Contributions by Treatment

Moreover, we find no relationship between *individual* RF behavior and contributions to the public good in either of the treatments without sorting. Pooling data from the NoSort- and Reverse-PG treatments, we estimate a mixed-effects panel regression of individual contributions on RF task waiting time, a period trend, and an intercept. We include random effects for each group and each individual-in-group to control for repeated measures, and we cluster standard errors at the group level. The regression, reported in Table E4 in the Appendix, reveals no significant relationship between individual RF behavior and contributions, though the period term is negative and significant. Therefore, we conclude that the sorting procedure

in the PG treatment eliminates cooperative decay in rule-following groups and that

the RF task does not measure other-regarding preferences *per se.*[22]

We interpret these results as evidence for a norm of *conditional cooperation.*

This interpretation is supported directly by additional mixed-effects panel regression

analysis that allows us to measure the conditionality of each individual's

contributions. Again restricting attention to the NoSort and Reverse treatments, we

regress individual contributions on a constant term, a period trend, the total time

the individual spent waiting in the RF task, the one-period lagged average of others'

contributions to the public good, a reverse treatment dummy, and an interaction

between waiting time and lagged others' contributions. We include random effects

for each group and subject-in-group to control for repeated measures, and we cluster

standard errors at the group level. Our estimates of the conditionality of cooperation

are contained in the coefficient on lagged others' contributions and its interaction

with RF task waiting time.

PG Contribution	Coef.	Std.Err.	z-value	Pr($> \|z\|$)
(Intercept)	36.288	4.797	7.56	0.000***
Period	−1.377	0.310	−4.45	0.000***
Reverse	−1.992	3.323	−0.60	0.549
Time Waited	−0.259	0.167	−1.55	0.121
Mean Others' Contributions$_{t-1}$	−0.029	0.120	−0.24	0.807
Time Waited×Mean Others' Contributions$_{t-1}$	0.009	0.005	1.76	0.078*

N = 1008; 28 groups × 4 subjects × 9 periods.
Standard errors clustered at the group level.
$* - (p < 0.10),^{**} -(p < 0.05),^{***} -(p < 0.01)$

TABLE 3. Mixed Effects Estimates of Contributions Contributions in PG, NoSort and Reverse Treatments

---

22. Non-parametric tests in Appendix E.2.1 also support this conclusion.

Regression output is reported in Table 3. A positive and marginally significant estimated coefficient on the interaction between lagged others' contributions and waiting time provides evidence of conditional cooperation among rule-followers – that is, contributions made by individuals who wait longer in the RF task are more responsive to the contributions made by others. Moreover, as we show in Appendix A, with our utility specification and under the assumption that such a norm exists, sustained cooperation is an equilibrium - but only when rule-followers are assortatively matched. The presence of rule-breakers, whose contributions naturally decay, breaks the cycle of conditional cooperation.

Finding 2: *When groups are not assortatively matched on RF task behavior, they exhibit cooperative decay. Individual RF behavior does not correlate with individual contributions, but rule-followers condition their contributions on the contributions of others.*

4.3. Trust Game Treatment

Recall that in the TG treatment, subjects are again assortatively matched. As in the PG treatment, there is considerable noise in group assignment as more than half of subjects followed the rule completely.

Figure 5 displays histograms of the amount sent by first movers from rule-following and rule-breaking groups. Mixed effects regression analysis cannot reject the null hypothesis of equal mean amount sent between the treatments (see Table E5 in Appendix E). However, the percent returned is higher in rule-following groups than
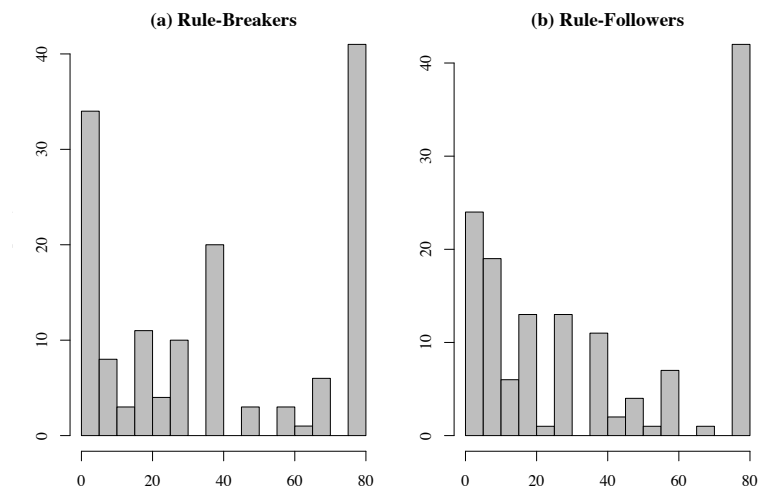
FIGURE 5. Histograms of Amount Sent in the TG treatment

in rule-breaking groups, as evidenced in Figure 6, which plots the average amount returned by second movers to first movers as a percent of the amount sent, for both group types in 3 bins.[23]

Statistical support is provided by additional mixed-effects regression analysis in Table 4. We regress the group average return on trust (amount returned/amount sent - 1) in each period on an intercept, a period trend, a rule-following group dummy and a period×rule-following interaction. A negative and significant intercept indicates that returns are negative in rule-breaking groups, while a positive and significant coefficient on the rule-following dummy shows that average returns (reciprocity) are higher among rule-followers. Wald tests confirm that rule-breaking groups provide

---

23.   In Appendix E, Figure E3 summarizes the full dataset from the TG treatment, showing, for each observation, the amount received by second-movers and the corresponding amounts returned and kept, by group type.
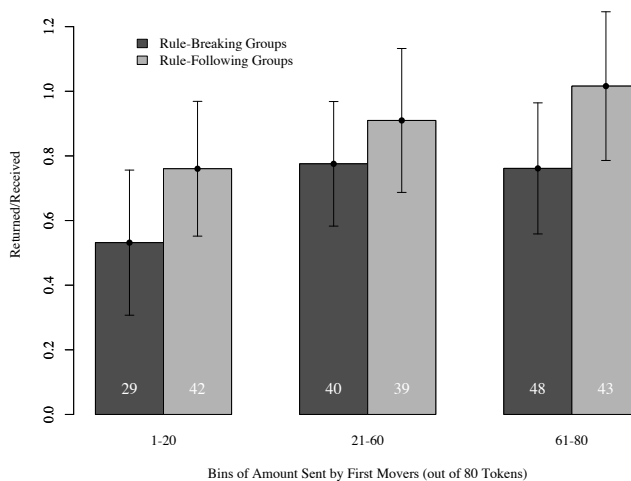
FIGURE 6. Barplots of *amount_returned/amount_sent* by receivers in the TG treatment for 3 bins of amount sent, ±2 SEs. The white number in each bar displays the # of observations in the bin for that group type (i.e. for followers and breakers).

significantly negative returns (p-value $= 0.006$), but rule-following groups provide returns statistically indistinguishable from zero (p-value $= 0.895$). Additional support for these conclusions, based on non-parametric tests can be found in Appendix E.

TG Return on Trust	Coef.	Std.Err.	z-value	Pr($> |z|$)
Intercept (Rule-Breakers)	-0.228	0.115	-1.97	0.049**
Period	-0.078	0.355	-2.18	0.029**
Rule-Following Group	0.364	0.200	1.82	0.068*
Rule-Following Group × Period	-0.076	0.073	-1.04	0.298

N = 72; 24 groups × 3 obs; standard errors clustered at the group level
$* - (p < 0.10),** -(p < 0.05),*** -(p < 0.01)$

TABLE 4. Mixed Effects Estimates of Returns on Trust by Group Type

Thus, as we show in Appendix A.1, our evidence is consistent with norm-dependent utility, assuming norms of both trust and reciprocity (or just of reciprocity). Crucially, even with a norm of trust, low-φ first-movers needn't send less than their high-φ counterparts as long as they believe that responders may be high-φ types who will follow the norm of reciprocity. Under this interpretation, at

least some giving in trust games may be interpreted as strategic investment on the part of first movers, while trustworthiness is rationalizable only among high-φ types who follow a norm of reciprocity.[24]

Finally, we note that our sample mean return is 25% of the amount received (75% of amount sent), and this is comparable to a previously observed mean of 20% for trust games in which subjects play both roles, reported in Table 1 of Johnson and Mislin (2011).

Finding 3: *Rule-followers exhibit more reciprocity than rule-breakers in the trust game, and there are no differences in the amount sent.*

4.4. Dictator Game Treatment

In the DG treatment, individuals were randomly assigned to roles, so we exploit the entire distribution of RF task behavior to compare dictator giving for various percentile cutoffs defining rule-following and rule-breaking. Overall, we observe a positive and marginally significant correlation between waiting time and amount sent (Spearman's $\rho = 0.191$, p-value $= 0.061$). However, as noted above, there is considerable measurement noise in the RF task, so it is reasonable to expect the effect size and significance to increase as we move further into the tails. With this fact in mind, panel (a) of Figure 7 displays mean amount sent by dictators of each

---

24.   If there is *no* norm of trust (or giving), then in our framework all TG giving is strategic. We return to the question of which norm is operant when we report our norm elicitation experiments in Section 4.6.

type, where the type is defined by a percentile cutoff in the RF-task distribution. As we move right along the x-axis, we are moving further into the tails of the RF-task distribution; the final data points, comparing the top and bottom deciles of the distribution, contain 7 observations of each type.
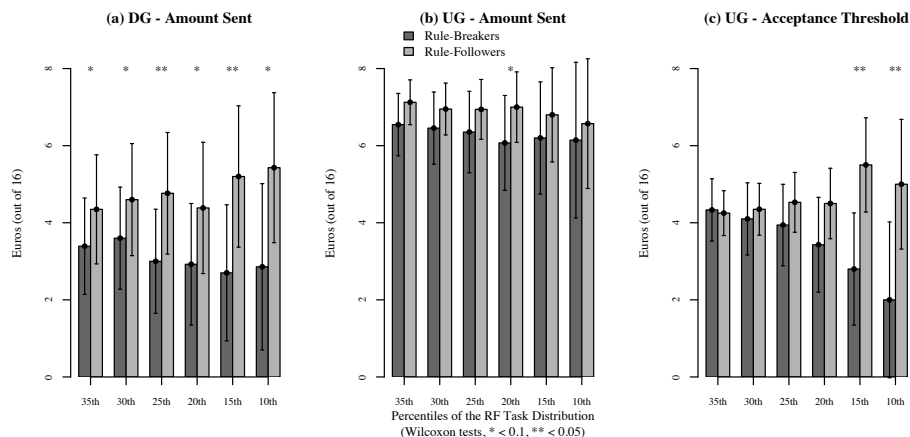

FIGURE 7. Behavior of Rule-Followers and Rule-Breakers in the DG and UG treatments. Each panel of the figure displays data relevant to one player role in one treatment, highlighting the comparison in behavior between rule-followers and rule-breakers. Panel (a) reports mean amounts sent by dictators in the DG treatment. Panel (b) reports mean amounts offered by proposers in the UG treatment. Panel (c) reports mean acceptance thresholds by responders in the UG treatment. The x-axis plots various percentile cutoffs defining rule-followers and rule-breakers; hence at the point labeled 25th, we are comparing means for the upper and lower quartiles of the RF task distribution (17 individuals per type). At the point labeled 10th, we are comparing deciles (7 individuals per type). Bars indicate ± 2 SEs. Stars indicate that rule-follower means are significantly greater than rule-breaker means by a one-sided Wilcoxon test. $* - (p < 0.1)$ and $** - (p < 0.05)$.

At each reported cutoff, a Wilcoxon test rejects the null hypothesis of equal mean amount sent by rule-followers and rule-breakers in favor of the alternative that rule-followers are more generous, and the difference grows as we move into the tails. Assuming a norm of sharing, this finding is consistent with the idea that the RF task measures the disutility of violating norms; see Section A.1 in the appendix. When comparing the upper and lower decile of the distribution, rule-followers (so-defined)

give nearly twice as much as rule-breakers. Moreover, if we classify as rule-followers all those who waited for at least 25 seconds in the RF task and as rule-breakers those who waited less than 25 seconds, our results remain essentially the same (p-value = 0.075).

Finally, we note that our pooled sample of 67 dictators yields an average amount sent of €4.01 or 25% of the total endowment, which is comparable to the mean of 28% reported in a meta-study of dictator games (Engel 2011).

Finding 4: *Rule-followers are more generous in the dictator game than rule-breakers.*

4.5. Ultimatum Game Treatment

Overall, we observe a positive but insignificant correlation between waiting time and amount sent (Spearman's $\rho = 0.113$, p-value = 0.179) and between waiting time and acceptance threshold ($\rho = 0.032$, p-value = 0.397). As in the DG treatment, we then classify rule-followers and rule-breakers according to their behavior in the RF task, and we offer statistical comparisons across multiple percentile cutoffs. Again, due to the noise in our measure of rule-following (our proxy for φ), we would expect both the magnitude and significance of any observed differences to increase as we move into the tails of the RF task distribution. Panel (b) of Figure 7 displays the mean amount offered by rule-followers and rule-breakers in the UG treatment. Panel (c) displays mean acceptance thresholds.

We observe no significant differences between rule-followers and rule-breakers in amount sent (with the exception of a weakly significant effect at the quintile cutoff),

though the mean is consistently €0.5 - €1 higher among rule-followers. Moreover, giving is notably higher among UG proposers than among DG proposers; rule-breaking proposers in the UG give more than even the most rule-following types in the DG. This is consistent with our framework if there exists a norm of sharing (and of rejecting unequal offers) but first movers are uncertain about the type of second movers. In such a case, many rule-breakers will nevertheless find it in their interest to send high amounts in order to avoid rejection; see Section A.1 in the appendix. Under this interpretation, our data provides evidence that giving in ultimatum games is a *strategic* decision, based in part on beliefs about the behavior of second-movers. Moreover, it provides further evidence that our RF task does not measure other-regarding distributional preferences *per se*, else we would expect similar proposer behavior in both the DG and UG.

Turning to second movers, as we move further into the tails of the RF-task distribution, the differences in acceptance threshold between rule-followers and rule-breakers become increasingly statistically and economically significant. For example, comparing the top and bottom decile of the waiting time distribution, rule-followers' mean threshold is €5 and rule-breakers' is €2. This difference is statistically significant by a one-sided Wilcoxon Test ($W_{7,7} = 38.5$, p-value $= 0.032$). Here behavior is consistent with the model if there exists a norm of rejecting unequal offers; see Section A.1 in the appendix.

As in the DG treatment, our pooled sample mean offer from 69 proposers is comparable to figures reported in the UG literature. Proposers offer 43% of the

endowment, compared to an average of 40% reported in a meta-study by Oosterbeek et al. (2004). Our pooled mean acceptance threshold set by responders is 27% of the endowment; unfortunately the meta-study does not report mean acceptance thresholds, though our observed threshold is comparable to the 33% reported in Table 2 of Schmitt et al. (2008).

Finding 5: *Rule-followers set higher acceptance thresholds than extreme rule-breakers in the ultimatum game, but there are no significant differences in offers. All types give more than in the dictator game.*

4.6. Elicited Norms

We have argued above and in Appendix A that our observations are rationalizable in a model of norm-dependent utility under certain reasonable assumptions about the character of the social norm in each game. To provide further empirical support for these arguments, we now report the results of additional experiments that allow us to measure subjects' beliefs about the social norm in each game (Section 3.5 describes these experiments in more detail). We present our findings in Figure 8 which shows subjects' beliefs about the modal subject's evaluations of each action's social appropriateness - i.e. beliefs about social norms in the PG, TG, DG and UG.

In the PG treatment, we elicited norms in the one-shot game as well as four hypothetical situations within the repeated game in which subjects had to evaluate potential actions *conditional* on past actions of others in the group. In Figure 8a we see that no matter the situation, subjects view high contributions as socially
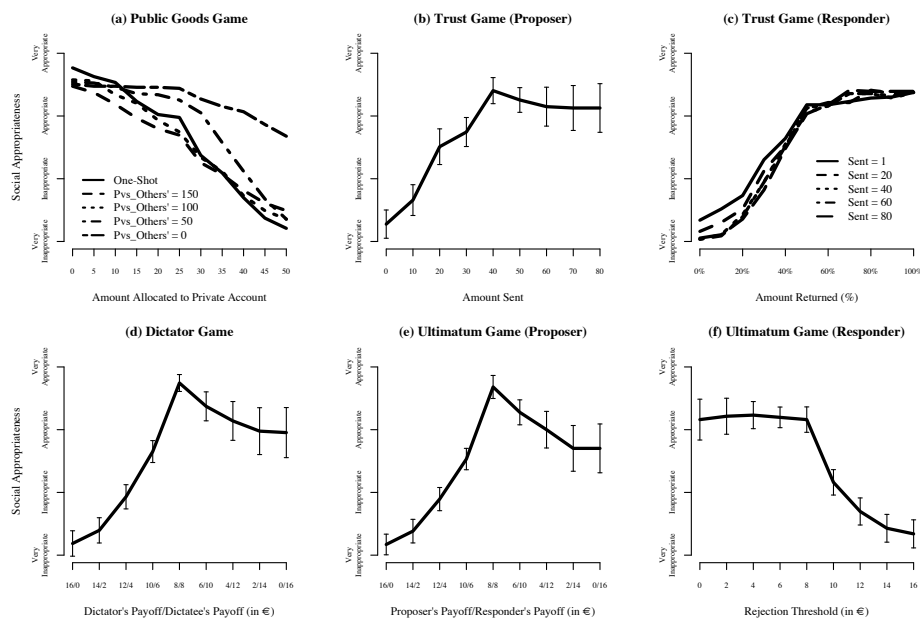
FIGURE 8. Elicited Social Norms. Points along the x-axis indicate the hypothetical action of the relevant individual. Lines denote mean "social appropriateness" of each action as elicited using the techniques due to Krupka and Weber (2013), while bars reveal ± 2 SEs. We suppress the bars in panels (a) and (c) to avoid cluttering the figure, which show elicitations for one player's actions *conditional* on actions by other players.

appropriate, but crucially as others' contributions at time $t-1$ decrease, subjects' ratings of selfish actions are increasingly positive, providing further evidence that a norm of conditional cooperation accounts for our observations in the PG.

Turning to the TG treatment in Figure 8b, there is no clear norm of either trust or altruism for proposers. In fact, the most "socially appropriate" action is to send half of the endowment, while there is little difference in the appropriateness of sending half and any greater amount (though the variance of ratings increases, indicating normative disagreement). Returning to Figure 5 we note that the amounts sent tend towards either extreme, even though sending 0 is the least appropriate action. Thus, giving in trust games is not driven by norms of *proposer* behavior. However,

Figure 8c reveals a norm of reciprocity for *responders* as the social appropriateness of returns increases sharply from 0% to the equal split and weakly from 50-100%. This is consistent with our observation that there is more reciprocity among groups of rule-followers.

In the DG treatment, reported in Figure 8d, there is a clear norm of equal division of the pie. Giving more than one-half of the pie tends to be viewed as more appropriate than giving less than one-half, though the high variance of those evaluations indicates considerable disagreement. Moreover, elicited norms in our dictator game are very similar to those in Krupka and Weber (2013). Under the elicited norm, observed DG behavior is consistent with our utility specification, providing further evidence for our claims in Section 4.4.

Finally, we examine norms in the UG. Proposer norms are shown in Figure 8e, and we observe normative beliefs very similar to those for proposers in the DG. There is a norm of equal division, and again, with high variance, more generous offers are rated as more appropriate than less generous offers. Thus, if proposer behavior were driven only by proposer norms, our findings that a) high- and low-φ types make roughly equal offers, and b) both types offer more than their counterparts in the DG, would appear inconsistent with the model. However, this ignores the strategic incentives faced by proposers who are aware of responder acceptance threshold norms.

Responder norms are reported in Figure 8f. Here, the most appropriate threshold is €4 (out of 16), though all thresholds up to and including the equal split are deemed roughly equally appropriate. Thus there is considerable heterogeneity in beliefs about

the appropriate threshold, and knowing this, low-φ proposers have incentive to make high offers in order to avoid rejection. This is consistent with our evidence on proposer behavior, and the considerable normative disagreement helps explain the relatively weaker results comparing rule-following and rule-breaking responders in the UG.

Taken together, elicited norms provide additional evidence in favor of our interpretation of sociality in our main treatments. Heterogeneity in norm-sensitivity (measured in the RF task) coupled with elicited social norms, jointly explain observed patterns of prosocial behavior and provide support for our contention that "norms make preferences social."

As an additional robustness check, in Appendix E.4 we report conditional logit regression analysis of decisions for each role in each game. We show that elicited norms interacted with our measure of φ have positive and significant effect on the probability of choosing an action in the DG and PG treatments. This indicates that subjects in these treatments with higher φ (as measured by our RF task) tend to put higher weight on the norm component of utility, further validating our framework. In the TG and UG treatments, the evidence is mixed. For proposers, a lack of significant effects is consistent with our interpretation that these decisions are strategic (and thus should not depend on own φ). For responders, insignificant effects in the UG are consistent with our weaker results there, in which only extreme rule-breakers exhibit large differences in rejection threshold. In the TG, the results point to negative weights on the norm and an offsetting positive interaction with assortative matching. This is partly because returning 100% of the amount received was rated as the most socially

appropriate action, but essentially no subjects returned more than 50%. It may also reflect the overall decay in reciprocity over time noted above.

Finding 6: *Elicited norms are largely consistent with our interpretation of the data in the PG, TG, DG and UG treatments.*

5. Conclusion

We explore a unifying framework for understanding prosociality in which individuals trade off own payoffs against a desire to adhere to social norms. After introducing a utility specification that captures this intuition, we develop a novel experiment that allows us to measure a proxy for the crucial parameter which reflects the extent to which an individual cares about norms. Using this information, we show that norm-following proclivity predicts play in public goods, trust, dictator, and ultimatum games, providing support for the idea that heterogeneous play in these games is driven by heterogeneous attitudes toward social norms. Additional experiments allow us to rule out explanations based on simple distributional preferences. We show how our main findings can be rationalized as a product of norm-dependent utility under reasonable assumptions about the relevant social norm, and we report additional experiments that directly elicit norms to provide further support for our interpretation. Taken together our data provide strong evidence for the idea that prosocial behavior in experiments is driven by prosocial norms - and not simply

by preferences for prosocial distributions of payoffs. Thus, norms make preferences (appear) social.

Taken in isolation, there are existing models of social preferences, reciprocity, and signaling that can account for each of our observations in the public goods, trust, dictator and ultimatum games. However, most of these models cannot account for all of these observations together, and moreover, they cannot explain the relationship between play in these games and behavior in the RF task. In our view, the reason these other models are sometimes successful in explaining the data is that they are all implicitly based on normative assumptions. When those normative assumptions match up with the perceived norms in the game, then the models do well. But this is also the reason that they are sometimes contradicted. When the implicit norm in the model doesn't match the norm perceived by the subjects, then the model fails. Thus, we shouldn't think of our results as a rejection of social preference models so much as a statement that each one reflects a special case of norm-dependence.

One avenue for future work will be to explore the robustness of our norm-sensitivity measurement task, in which subjects are instructed to follow a costly rule. One concern with the present version of the RF task is that subjects from different cultures may import different norms associated with traffic lights. If some of our measured rule-following results from these imported norms, this could be a source of noise in measuring φ. Indeed, the "NoRule" treatment, in which 12% of subjects still follow the rule provides some support for this. Nevertheless, since our data indicate that individuals are extremely responsive to a simple statement that "the rule is..."

it will be useful to develop context-minimal screening tasks to reduce noise in the proxy for φ.[25]

The most important unanswered question, though, and the one that we hope this research will encourage others to ask, is "where do norms come from?" For simplicity, we take them to be exogenous, but we could also think of them as being specific to an individual's identity, social group, or culture (this is one way to view Akerlof and Kranton 2000). To the extent that this is true, the same experimental procedures may induce entirely different norms, depending on the cultural background of the subjects. If the social norms associated with a particular context differ across cultures, then cross-cultural behavioral differences in laboratory experiments also come as no surprise – and we can explain these differences while maintaining that people facing different norms nevertheless have the same underlying motivations.[26]

As Wilson (2008) has argued, "in general, cooperative outcomes are the product of human agreement, tacit or otherwise, on the social context of the interaction" (p. 374). Since our experimental environment suppresses communication, any agreement on the norms of action is necessarily tacit, and the extent of tacit agreement is likely tied to

---

25. To this end, we have recently collected data in a simple variant of the RF task with this goal in mind. Subjects are instructed that "the rule is" to place virtual balls into one of two virtual buckets, where they earn a money return for each ball placed in each bucket. The alternative bucket provides twice the return of the "rule" bucket. RF behavior in this task is highly correlated with dictator giving (Spearman's $\rho = 0.41$, p-value < 0.01); see (Kimbrough et al. 2014).

26. See e.g. Roth et al. (1991); Henrich et al. (2005); Herrmann et al. (2008).

the extent to which subjects share a cultural/experiential background. Although our subject pool contains individuals from a large number of nations, the preponderance of our subjects hail from Western Europe and were raised according to the rules and norms common to European culture(s). This likely encouraged cooperation among our high-φ types by increasing agreement about the norm.

In a similar vein, Roth et al.'s (1991) findings that subject behavior differs across cultures in non-market contexts—but is essentially the same in market contexts—has interesting implications. One interpretation is that markets are norm-free, that is, behavior in markets is culturally invariant because markets work around or outside of normative concerns.[27] Another interpretation is that the norms associated with markets are common across cultures. This is likely true to some extent, though clearly there are cultural differences in the types of things that are viewed as commodities and the kinds of market transactions that are deemed acceptable (Roth 2007). However, both hypotheses are difficult to reconcile with the evidence in Henrich et. al (2010) that greater exposure to markets and to large-scale institutions such as organized religion are both correlated with experimental measures of other-regarding and cooperative behavior. Instead, one might argue that certain norms are *embedded* in market institutions, and they are transmitted through repeated interaction. We leave these questions for future research.

---

27. This is related to the argument in Fehr and Schmidt (1999), where the effects of social preferences in their model dissipate when individuals are small relative to the market; competition limits the effectiveness of prosocial action.

References

Akerlof, George A. and Rachel E. Kranton (2000). "Economics and Identity." *Quarterly Journal of Economics*, 115(3), 715–753.

Alger, Ingela and Jörgen Weibull (2013). "*Homo Moralis* - Preference Evolution under Incomplete Information and Assortative Matching." *Econometrica*, 81(6), 2269–2302.

Andreoni, James (1995). "Warm-Glow Versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments." *Quarterly Journal of Economics*, 110(1), 1–21.

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 John Miller (2002). "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, 70(2), 737–753.

Bardsley, Nicholas (2008). "Dictator game giving: altruism or artefact?" *Experimental Economics*, 11, 122–133.

Bénabou, Roland and Jean Tirole (2006). "Incentives and Prosocial Behavior." *The American Economic Review*, 96(5), 1652–1678.

Berg, Joyce, John Dickhaut, and Kevin McCabe (1995). "Trust, reciprocity, and social history." *Games and Economic Behavior*, 10(1), 122–142.

Bicchieri, Cristina (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bolton, Gary E. and Axel Ockenfels (2000). "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, 90(1), 166–193.

Brandts, Jordi and Gary Charness (2011). "The strategy versus the direct-response method: a first survey of experimental comparisons." *Experimental Economics*, 14, 375–398.

Burnham, Terence, Kevin McCabe, and Vernon L. Smith (2000). "Friend-or-foe intentionality priming in an extensive form trust game." *Journal of Economic Behavior & Organization*, 43, 57–73.

Cappelen, Alexander W., Astri Drange Hole, Erik Ø. Sørensen, and Bertil Tungodden (2007). "The Pluralism of Fairness Ideals: An Experimental Approach." *American Economic Review*, 97(3), 818–827.

Charness, Gary and Matthew Rabin (2002). "Understanding Social Preferences With Simple Tests." *Quarterly Journal of Economics*, 117(3), 817–869.

Charness, Gary and Arthur Schram (2013). "Social and Moral Norms in Allocation Choices in the Laboratory." Mimeo, University of California at Santa Barbara and University of Amsterdam.

Cherry, Todd L., Peter Frykblom, and Jason F. Shogren (2002). "Hardnose the Dictator." *American Economic Review*, 92(4), 1218–1221.

Cialdini, Robert B. and Melanie R. Trost (2008). "Social Influence: Social Norms, Conformity, and Compliance." In *The Handbook of Social Psychology*, vol. 1-2, edited by Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, pp. 151–192. McGraw-Hill, New York, NY.

Cox, James C., Daniel Friedman, and Steven D. Gjerstad (2007). "A Tractable Model of Reciprocity and Fairness." *Games and Economic Behavior*, 59(1), 17–45.

Cox, James C., Daniel Friedman, and Vjollca Sadiraj (2008). "Revealed Altruism." *Econometrica*, 76(1), 31–69.

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

Dufwenberg, Martin and Georg Kirchsteiger (2004). "A theory of sequential reciprocity." *Games and Economic Behavior*, 47, 268–298.

Eckel, Catherine and Philip Grossman (1996). "Altruism in anonymous dictator games." *Games and Economic Behavior*, 16, 181–191.

Elster, Jon (1989). "Social norms and economic theory." *The Journal of Economic Perspectives*, 3(4), 99–117.

Engel, Christoph (2011). "Dictator games: A meta study." *Experimental Economics*, 14(4), 583–610.

Falk, Armin and Urs Fischbacher (2006). "A theory of reciprocity." *Games and Economic Behavior*, 54, 293–315.

Fehr, Ernst and Simon Gächter (2000). "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14(3), 159–181.

Fehr, Ernst and Klaus M. Schmidt (1999). "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114(3), 817–868.

Fershtman, Chaim, Uri Gneezy, and John A. List (2012). "Equity aversion: Social norms and the desire to be ahead." *American Economic Journal: Microeconomics*, 4(4), 131–144.

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

Fisman, Raymond, Shachar Kariv, and Daniel Markovits (2007). "Individual Preferences for Giving." *American Economic Review*, 97(5), 1858–1876.

Forsythe, Robert, Joel L. Horowitz, Nathan E. Savin, and Martin Sefton (1994). "Fairness in simple bargaining experiments." *Games and Economic behavior*, 6(3), 347–369.

Goeree, Jacob K. and Charles A. Holt (2001). "Ten little treasures of game theory and ten intuitive contradictions." *American Economic Review*, 91(5), 1402–1422.

Graham, Jesse, Jonathan Haidt, and Brian Nosek (2008). *The Moral Foundations Quiz.* www.yourmorals.org.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982). "An experimental analysis of ultimatum bargaining." *Journal of Economic Behavior & Organization*, 3(4), 367–388.

Haidt, Jonathan and Craig Joseph (2004). "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues." *Daedalus*, 133(4), 55–66.

Hayek, Friedrich A. (1973). *Law, Legislation and Liberty, vol. 1: Rules and Order.* The University of Chicago Press.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, Michael Alvard, Abigail Barr, Jean Ensminger, Natalie Smith Henrich, Kim Hill, Francisco Gil-White, Michael Gurven, Frank W. Marlowe, John Q. Patton, and

David Tracer (2005). ""Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies." *Behavioral and Brain Sciences*, 28, 795–855.

Henrich, Joseph, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker (2010). "Markets, religion, community size, and the evolution of fairness and punishment." *Science*, 327(5972), 1480–1484.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter (2008). "Antisocial Punishment Across Societies." *Science*, 319, 1362–1367.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon L. Smith (1994). "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 7(3), 346–380.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith (1996). "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review*, 86(3), 653–60.

Hoffman, Elizabeth and Matthew L. Spitzer (1985). "Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice." *Journal of Legal Studies*, 14, 259–298.

Isaac, R. Mark and James M. Walker (1988). "Group size effects in public goods provision: The voluntary contributions mechanism." *Quarterly Journal of Economics*, 103(1), 179–199.

Johnson, Noel D. and Alexandra A. Mislin (2011). "Trust games: A meta-analysis." *Journal of Economic Psychology*, 32(5), 865–889.

Kessler, Judd B. and Steven Leider (2012). "Norms and Contracting." *Management Science*, 58(1), 62–77.

Kimbrough, Erik O., Joshua B. Miller, and Alexander Vostroknutov (2014). "Norms, Frames and Prosocial Behavior in Games." Mimeo, Simon Fraser University, Bocconi University, Maastricht University.

Krupka, Erin L. and Roberto A. Weber (2009). "The focusing and informational effects of norms on pro-social behavior." *Journal of Economic Psychology*, 30(3), 307–320.

Krupka, Erin L. and Roberto A. Weber (2013). "Identifying social norms using coordination games: Why does dictator game sharing vary?" *Journal of the European Economic Association*, 11(3), 495–524.

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), 136–163.

Levine, David K. (1998). "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1(3), 593–622.

Levitt, Steven D. and John A. List (2007). "What do laboratory experiments measuring social preferences reveal about the real world?" *The Journal of Economic Perspectives*, 21(2), 153–174.

Lewis, David K. (2002). *Convention: A Philosophical Study*. Blackwell, New York.

List, John A. (2007). "On the Interpretation of Giving in Dictator Games." *Journal of Political Economy*, 115(3), 482–493.

López-Pérez, Raúl (2008). "Aversion to norm-breaking: A model." *Games and Economic behavior*, 64(1), 237–267.

McCabe, Kevin, Mary L. Rigdon, and Vernon L. Smith (2003). "Positive reciprocity and intentions in trust games." *Journal of Economic Behavior and Organization*, 52, 267–275.

Merton, Robert K. (1957). *Social Theory and Social Structure*. Free Press.

Oosterbeek, Hessel, Randolph Sloof, and Gijs van de Kuilen (2004). "Cultural differences in ultimatum game experiments: Evidence from a meta-analysis." *Experimental Economics*, 7(2), 171–188.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/. ISBN 3-900051-07-0.

Rabin, Matthew (1993). "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83(5), 1281–1302.

Roth, Alvin E. (2007). "Repugnance as a Constraint on Markets." *The Journal of Economic Perspectives*, 21(3), 37–58.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir (1991). "Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study." *American Economic Review*, 81(5), 1068–1095.

Schmitt, Pamela, Robert Shupp, Kurtis Swope, and Justin Mayer (2008). "Pre-commitment and personality: Behavioral explanations in ultimatum games." *Journal of Economic Behavior & Organization*, 66(3), 597–605.

Sherif, Muzafer (1936). *The Psychology of Social Norms*. Harper.

Smith, Adam (1759). *The Theory of Moral Sentiments*. Liberty Fund: Indianapolis (1982).

Smith, Vernon L. (2008). *Rationality in Economics: Constructivist and Ecological Forms*. New York, Cambridge University Press.

Smith, Vernon L. and Bart J. Wilson (Forthcoming). "*Sentiments*, Conduct and Trust in the Laboratory." *Social Philosophy and Policy*.

Wilson, Bart J. (2008). "Language games of reciprocity." *Journal of Economic Behavior & Organization*, 68(2), 365–377.

Supporting Information

Additional Supporting Information including appendices may be found in the online version of this article on the publisher's website.