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)

The editor in charge of this paper was Stefano DellaVigna.

Acknowledgments: This paper benefited immeasurably (title included) from repeated discussions with Ryan Oprea. We also thank Jeff Butler, Luke Coffman, Yoram Halevy, Dan Houser, Taylor Jaworski, Judd Kessler, Erin Krupka, Krishna Pendakur, Arno Riedl, Jared Rubin, Vernon Smith, Bart Wilson, participants in seminars at the University of British Columbia, University of Arkansas, University of Alabama, Concordia University, Simon Fraser University, and conference participants at the 2011 North American meetings of the ESA, the 2012 meetings of the Association for the Study for Religion, Economics, and Culture, and at the 2013 North American meetings of the ESA for helpful comments. We also gratefully acknowledge funding from Maastricht University's METEOR research school, the European Union Marie Curie FP7 grant program, and the Social Sciences and Humanities Research Council of Canada's Insight Development Grants Program. Finally, we would like to thank two referees and the editor for their detailed comments which improved the paper immensely. The data are available according to the policy of the journal. Any remaining errors are our own.

E-mail: ekimbrou@sfu.ca (Kimbrough); a.vostroknutov79@gmail.com (Vostroknutov)

Journal of the European Economic Association June 2016 14(3):608–638
 DOI: 10.1111/jeea.12152

1. Introduction

Over the last 30 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 preferences over payoff distributions, there is evidence that many people trade off own and others' 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, for example 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-signaling (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 reputations are tied to their degree of prosociality, and they account for many of the observations noted previously 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 (Cappelen et al. 2007; López-Pérez 2008; Krupka and

1. See, for example, 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), and Cox et al. (2007, 2008).

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), and Smith and Wilson (forthcoming).

Weber 2009, 2013; Kessler and Leider 2012).[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 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 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.

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.

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—that is, 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, p. 824) put it, "any reciprocal model must embed assumptions about distributional preferences", 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.

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 voluntary contribution mechanism (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. Because 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 previously, 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? 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]

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

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 previously). Briefly, norm-dependence implies a utility function that is increasing in own payoff and decreasing in the deviation between own action and the normative action. We introduce a norm-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 Online 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 everything (as might be the case when the proposer has *earned* the right to allocate the pie; see for example 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 $U_p(x, A) = x - \varphi_p g(|x - 1|)$

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, which follows Krupka and Weber 2013).

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 Online 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 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 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 Online 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 RF 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 in what follows.

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 five decisions concerning how long they wait at a sequence of red traffic lights, each of which will turn green five seconds after their arrival. Figure 1 shows the screen that subjects see.

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).

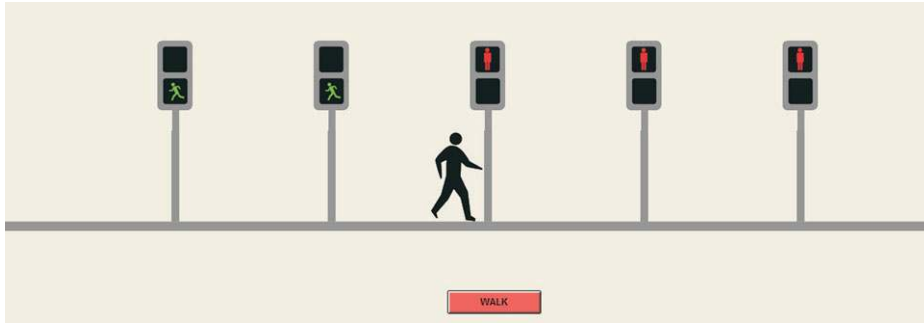


FIGURE 1. Screen shot of the RF task.

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 five 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 four seconds to walk between each light, and four seconds from the final light to the finish. Therefore, all subjects lose around €2 walking, and if a subject waits for green at all five 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 five seconds at each light). In the instructions for the RF task (see Online 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 RF 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

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.

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.

drivers and pedestrians—has no significance because there are no other drivers or pedestrians to protect or be protected from. Yet in our experience, 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, because all costs and benefits are private in the RF task, social preference 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 ten periods of a repeated public goods game with a VCM in fixed groups of four (Isaac and Walker 1988). Each subject receives an endowment of 50 tokens at the beginning of each period (one 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 $0.5 \times$ (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 eight (sessions consisted of 16, 24, or 32 subjects). Second, within each group of eight, 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

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 Online 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.

close to 0 seconds for those who did not wait at any light. Then, in each group of eight, we separate the top four subjects (rule-followers) and the bottom four subjects (rule-breakers) into two groups for Stage 2. After we match subjects, there is no interaction between any groups of four. 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 Online 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 four on the basis of their RF task behavior, and then they play the trust game six 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 three times in the role of first mover (blue person) and three times as a second mover (red person; see Online 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 (one 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 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 Online 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

14. Note that we did not deceive our subjects. None of the statements in the instructions is false or misleading. It is a separate, and also interesting, question whether subject's 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.

15. In a meta-study, Brandts and Charness (2011) find that in experiments involving punishment or rejection, the strategy method generally gives different results than the standard extensive-form approach. Nevertheless, we are making comparisons between rule-followers and rule-breakers who all used the strategy method, and we elicit norms in Section 4.6 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.

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.

3.2.4. Diagnostic Treatments. As robustness checks meant to rule out some alternative hypotheses, we also ran the following diagnostic treatments, which we 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 three 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 "five seconds after the stick figure reaches a stop light, it will turn from red to green";
4. a NoRule-Reverse-PG treatment combining treatments 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, this allows us to better understand the norms underlying behavior by observing mixed-type groups. The NoRule treatments, however, allow us to determine what portion of observed rule-following is due to 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

16. This is designed to measure the strength of subjects' respect for various moral values (Haidt and Joseph 2004; Graham et al. 2008; see Online 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 six 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.

TABLE 1. Summary of main experimental design.

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
No. of subjects	72	64	24	48	24	96	134	138
No. of groups	18	16	6	12	6	24	67	69
No. of obs. per group	10	10	10	10	10	6	1	1

Notes: RF, rule-following task; NoRule, RF task with no rule; DG, dictator game; UG, ultimatum game; TG, trust game; PG, voluntary contributions PG.

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 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 Online 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 Online 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 nonzero 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. Online 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 in what follows 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 are 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, because 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 2(a) displays a histogram of waiting times in the RF task. The 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.

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

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.

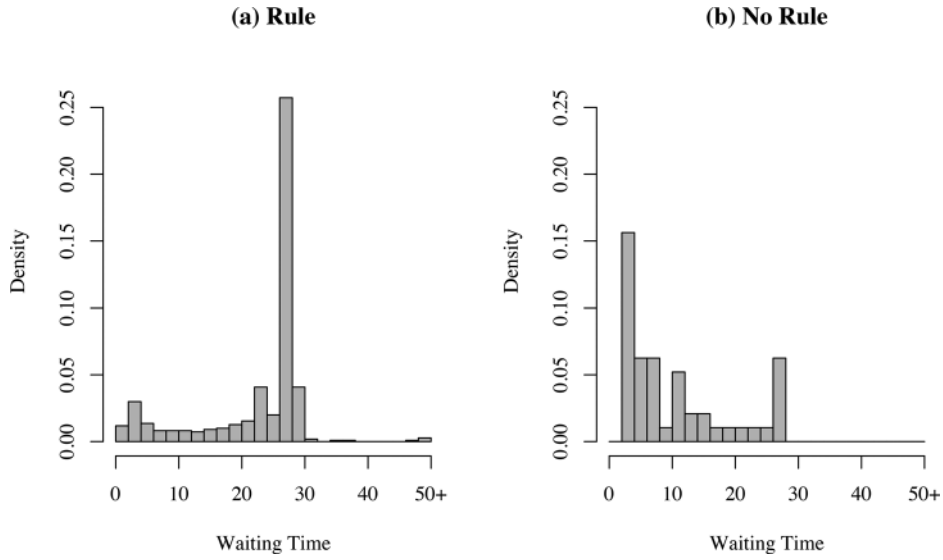


FIGURE 2. Histograms of waiting time in seconds, Rule versus No-Rule treatments.

this reason we conducted an additional "NoRule" treatment—described in Section 3.2. Figure 2(b) 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, the 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 Online 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 eight, and then those groups were split into "rule-following" and "rule-breaking" groups at the within-group median waiting time. The mean waiting time in rule-breaking groups is 19 seconds (SE $= 0.98$), while the mean waiting time in rule-following groups is 27 seconds (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 four lights.[19] This suggests

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

19. The relatively high waiting times among rule-followers (> 27 seconds) can be explained by reaction times—a lag of ∼0.5 seconds between the light changing and clicking WALK.

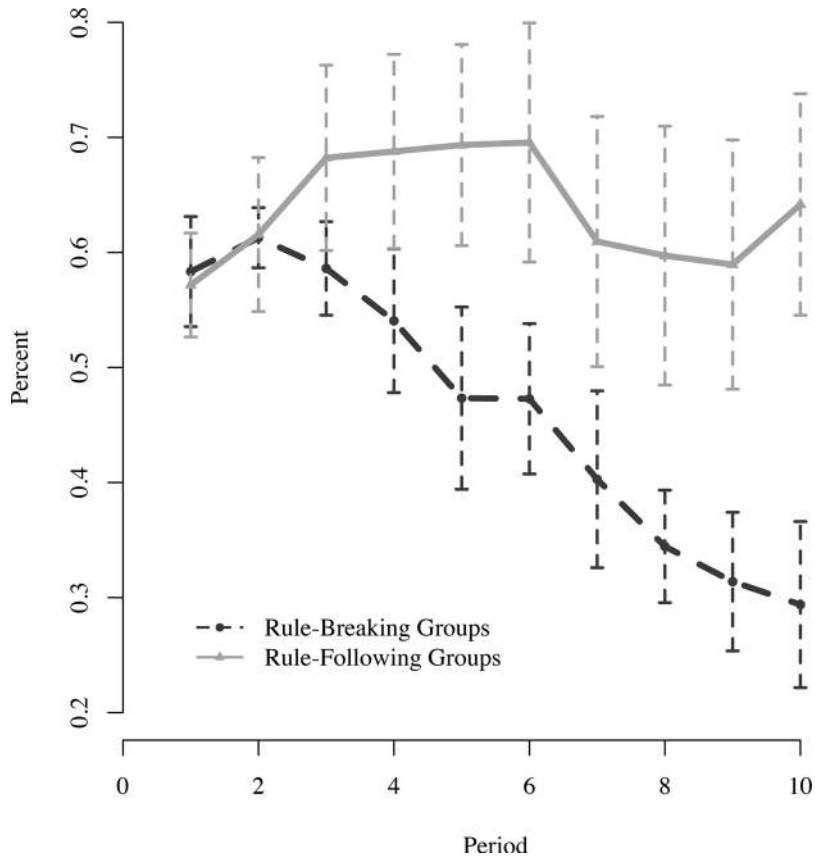


FIGURE 3. Time series of mean percent of endowment contributed ±1 SE, for rule-following and rule-breaking groups in the PG treatment (computed at the group level; nine independent observations underlying each line).

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.

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

TABLE 2. Mixed-effects estimates of group mean PG contributions (out of 50 tokens).

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 × period	1.851	0.613	3.02	0.003***

Notes: $N = 180$; 18 groups × 10 periods; standard errors clustered at the group level.
***Significant at 1%.

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 significant coefficient on the period × 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 Online Appendix E.2.1 we provide nonparametric 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 Online 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]

20. If we also include the six groups from our No-Rule treatment in the analysis, which were also sorted according to waiting time, the results remain essentially unchanged. Figure E.2 in Online Appendix E shows time series including sorted NoRule sessions in the computation of rule-following and rule-breaking means.

21. This model can also help to 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.

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

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 four in the NoSort-PG treatment and 12 groups of four 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 E.3 in Online Appendix E supports the evidence presented in Figure 4.

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 E.4 in the Online 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

22. Nonparametric tests in Online Appendix E.2.1 also support this conclusion.

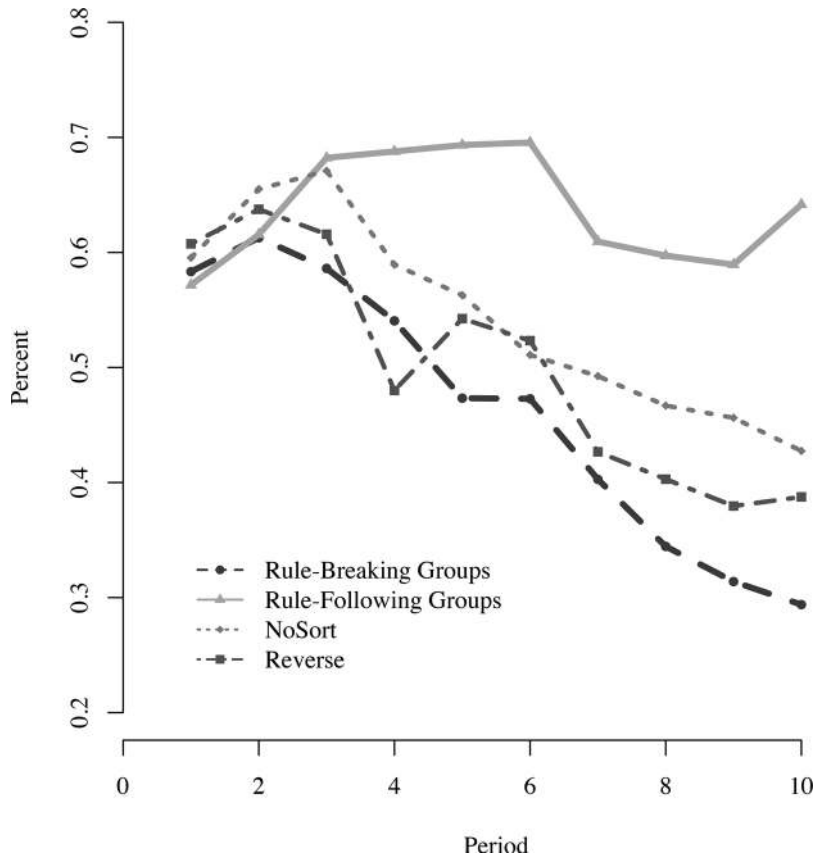


FIGURE 4. Time series of mean group public good contributions by treatment.

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.

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 Online 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.

TABLE 3. Mixed-effects estimates of contributions in PG, NoSort, and Reverse treatments.

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*

Notes: $N = 1008$; 28 groups × 4 subjects × 9 periods. Standard errors clustered at the group level.
*Significant at 10%; ***significant at 1%.

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

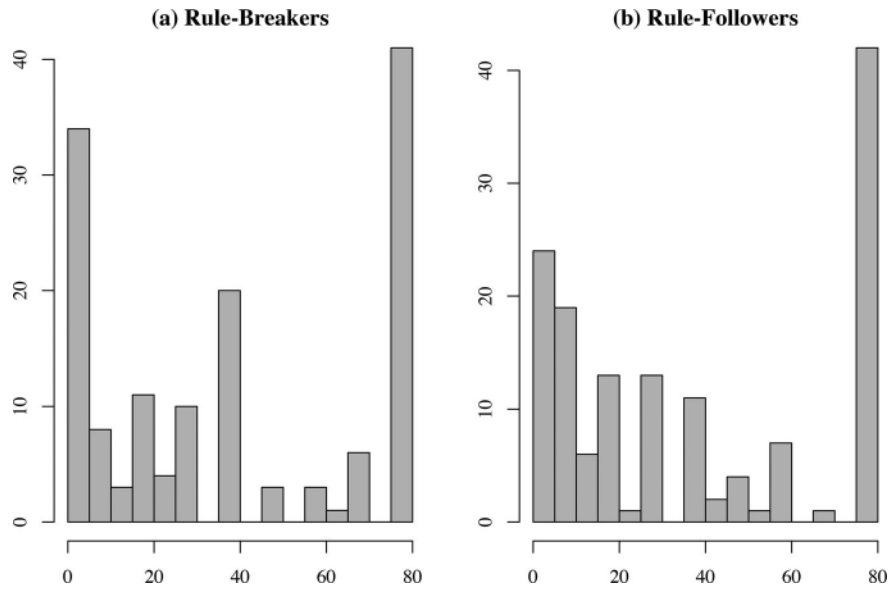


FIGURE 5. Histograms of amount sent in the TG treatment.

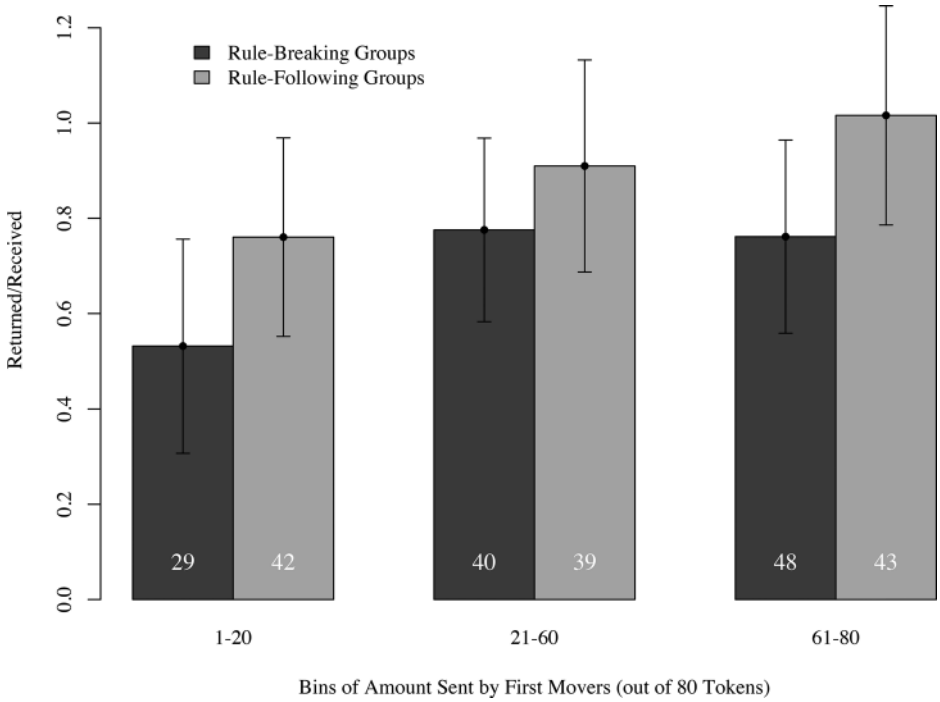


FIGURE 6. Barplots of amount returned/amount sent by receivers in the TG treatment for three bins of amount sent, ±2 SEs. The white number in each bar displays the number of observations in the bin for that group type (i.e., for followers and breakers).

hypothesis of equal mean amount sent between the treatments (see Table E.5 in Online Appendix E). However, the percentage returned is higher in rule-following groups than in rule-breaking groups, as evidenced in Figure 6, which plots the average amount returned by second movers to first movers as a percentage of the amount sent, for both group types in three 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 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 nonparametric tests can be found in Online Appendix E.

23. In Online Appendix E, Figure E.3 summarizes the full data set from the TG treatment, showing, for each observation, the amount received by second-movers and the corresponding amounts returned and kept, by group type.

TABLE 4. Mixed-effects estimates of returns on trust by group type.

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

Notes: $N = 72$; 24 groups × 3 observations. Standard errors clustered at the group level.
*Significant at 10%; **significant at 5%.

Thus, as we show in Online 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 need not 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 previously, 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, Figure 7(a) displays the mean amount sent by dictators of each 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 seven observations of each type.

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.

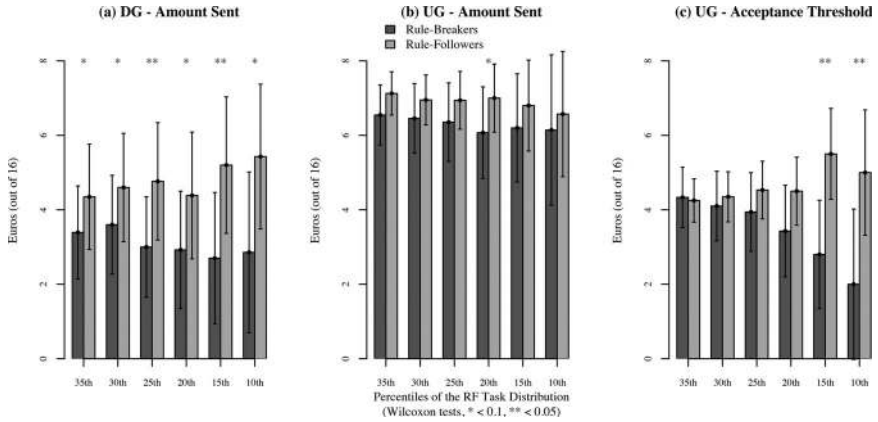


FIGURE 7. Behavior of rule-followers and rule-breakers in the DG and UG treatments. Each panel of the figure displays data relevant to one player role in one treatment, highlighting the comparison in behavior between rule-followers and rule-breakers. Panel (a) reports mean amounts sent by dictators in 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. *Significant at 10%; **significant at 5%.

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 Online Appendix A.1. 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. Figure 7(b) displays the mean amount offered by rule-followers and rule-breakers in the UG treatment. Figure 7(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 extreme 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 Online Appendix A.1. Under this interpretation, our data provide 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 deciles of the waiting time distribution, the rule-followers' mean threshold is €5 and the 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 Online Appendix A.1.

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 previously and in Online Appendix A that our observations are rationalizable in a model of norm-dependent utility under certain reasonable

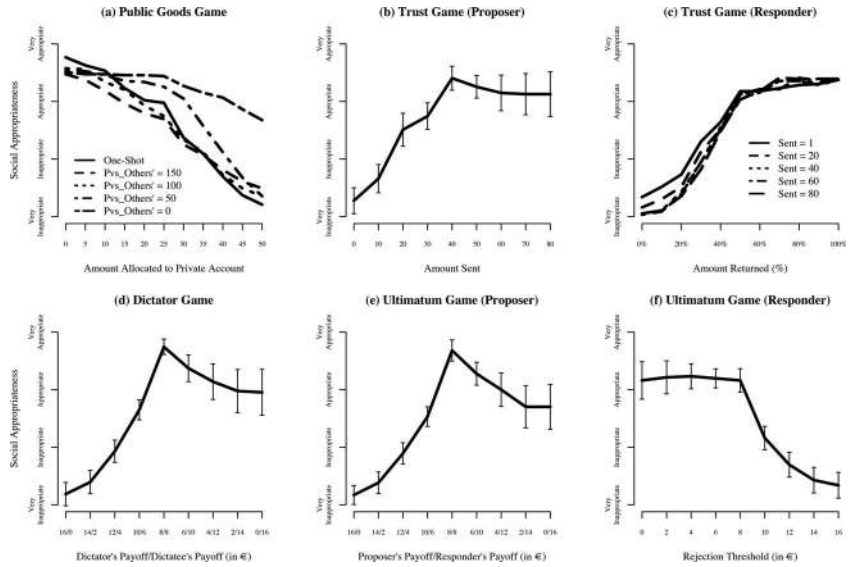


FIGURE 8. Elicited social norms. Points along the x-axis indicate the hypothetical action of the relevant individual. Lines denote mean "social appropriateness" of each action as elicited using the techniques due to Krupka and Weber (2013), while bars reveal ± 2 SEs. We suppress the bars in panels (a) and (c) to avoid cluttering the figure, which show elicitations for one player's actions *conditional* on actions by other players.

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—that is, 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 8(a), we see that no matter the situation, subjects view high contributions as socially 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 8(b), 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 8(c) 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 8(d), 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 8(e), 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 8(f). 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 Online 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 previously.

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. However, this is also the reason that they are sometimes contradicted. When the implicit norm in the model does not match the norm perceived by the subjects, then the model fails. Thus, we should not 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, because 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, p. 374) has argued, "in general, cooperative outcomes are the product of human agreement, tacit or otherwise, on the social context of the interaction". Because 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 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, the finding of Roth et al. (1991) that subject behavior differs across cultures in nonmarket 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.

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, for example, Roth et al. (1991), Henrich et al. (2005), and Herrmann et al. (2008).

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, 715–753.

Alger, Ingela and Jörgen Weibull (2013). "Homo Moralis—Preference Evolution under Incomplete Information and Assortative Matching." *Econometrica*, 81, 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, 1–21.

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

Andreoni, James and John Miller (2002). "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, 70, 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." *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, 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 and 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, 817–869.

Charness, Gary and Arthur Schram (2013). "Social and Moral Norms in Allocation Choices in the Laboratory." Working paper, 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*. edited by D. T. Gilbert, S. T. Fiske, and G. Lindzey. McGraw-Hill, Vol. 1–2, pp. 151–192.

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

Cox, James C., Daniel Friedman, and Vjollca Sadiraj (2008). "Revealed Altruism." *Econometrica*, 76, 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, 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." *Journal of Economic Perspectives*, 3(4), 99–117.

Engel, Christoph (2011). "Dictator Games: A Meta Study." *Experimental Economics*, 14, 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, 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, 131–144.

Fischbacher, Urs (2007). "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics*, 10, 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, 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 and Organization*, 3, 367–388.

Haidt, Jonathan and Craig Joseph (2004). "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues." *Daedalus*, 133, 55–66.

Hayek, Friedrich A. (1973). *Law, Legislation and Liberty, Vol. 1: Rules and Order*. 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, 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, 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), 653–660.

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, 179–199.

Johnson, Noel D. and Alexandra A. Mislin (2011). "Trust Games: A Meta-Analysis." *Journal of Economic Psychology*, 32, 865–889.

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

Kimbrough, Erik O., Joshua B. Miller, and Alexander Vostroknutov (2014). "Norms, Frames and Prosocial Behavior in Games." Working paper, 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, 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, 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, 136–163.

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

Levitt, Steven D. and John A. List (2007). "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives*, 21(2), 153–174.

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

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

López-Pérez, Raúl (2008). "Aversion to Norm-Breaking: A Model." *Games and Economic Behavior*, 64, 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, 171–188.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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." *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 and Organization*, 66, 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*. 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 and Organization*, 68, 365–377.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Online Appendix

Data and do files

ONLINE APPENDIX FOR "NORMS MAKE PREFERENCES SOCIAL"

Erik O. Kimbrough
Simon Fraser University

Alexander Vostroknutov
Maastricht University

The editor in charge of this paper was Stefano DellaVigna.

E-mail: ekimbrou@sfu.ca (Kimbrough); a.vostroknutov79@gmail.com (Vostroknutov)

Journal of the European Economic Association
Preprint prepared on 5 May 2015 using jeea.cls *v1.0.*

Erik O. Kimbrough and Alexander Vostroknutov

Appendices: Norms Make Preferences Social

A Norm-Dependent Utility

The goal of this section is to report a simple model in which we can illustrate how the notion of norm-dependent utility - the idea that utility is defined over both own payoffs and own norm adherence - combined with suitably defined and commonly shared social norms can account for heterogeneity in prosocial behavior in our experiments. Moreover, we highlight throughout other stylized facts from the experimental literature that are also potentially explicable under norm-dependent utility. We do not mean for our particular formulation to be the last word on modeling the role of norms in economic behavior; instead, we use our model to ground our interpretation of the data.

A.1 Norm-Dependence in Some Games

In all three models of games introduced below we maintain the same definitions: $g : [0,1] \to [0,1]$ is a strictly convex increasing differentiable function with $g(0) = 0$ and $g(1) = 1$, which represents the disutility of deviation from the norm; $\phi_p, \phi_r \geq 0$ is the norm sensitivity parameter of proposer and responder (in Trust and Ultimatum games).

Dictator Game. Let the choice of the proposer in the dictator game be $x \in [0,2]$ and assume that the norm prescribes equal division of the pie. Then the proposer's norm-dependent utility can be defined as

$$U_p(x) = x - \phi_p g(|x-1|).$$

Thus, following the norm gives no disutility to the proposer, while choosing to keep everything ($x = 2$) gives the maximal disutility of ϕ_p.

Given the assumptions on g, the optimal choice $x^*(\phi_p) \in [1,2]$ of the proposer is weakly decreasing in ϕ_p. For $\phi_p \to 0$, proposer chooses $x^* \to 2$; for $\phi_p \to \infty$ she chooses $x^* \to 1$. Since g is assumed strictly convex, all intermediate values $x^*(\phi_p) \in (1,2)$ are possible for some ϕ_p. Note also that the evidence that people are sensitive to the "price" of giving (reported in Andreoni and Miller, 2002; Fisman et al., 2007) is consistent with the tradeoffs implied by our model.

We could also consider another norm governing choice in the dictator game. Suppose that the dictator first earned the right to allocate the money through some pre-play task and then faces the decision in the dictator game. By introducing competition to assign "property rights," the normative action changes (Bicchieri, 2006). Players who have earned the right to be dictator also believe they have earned a right to a larger share of the pie. In this case we can assume that $\eta = k > 1$. The utility of the proposer becomes $U_p(x) = x - \phi_p g(|x-k|/k)$, so that $x^*(\phi_p) \in [k,2]$ and the proposer will choose to keep a larger share of the money. Indeed, in a variety of laboratory experiments with earned rights, this is exactly what is observed (e.g Hoffman et al., 1996). Moreover, if $\eta = 2$, we get $x^*(\phi_p) = 2$ for all ϕ_p, or, in other words, the proposer will keep the entire pie as was observed in Cherry et al. (2002) who combined earned rights and double-blind protocols, sharply shifting social expectations towards selfishness.

Ultimatum Game. In the ultimatum game (Güth et al., 1982) the proposer chooses a division of the pie $x \in [0,2]$, where x is the amount she decides to keep for herself. Then, the responder decides to accept the

division $(x, 2-x)$ or to reject it, in which case both players get 0. Suppose that the norm prescribes the proposer to divide the pie equally and the responder to accept any offer $2-x \geq 1$, which gives him at least half, and reject all offers $2-x < 1$, which give the responder less than half. Then, the norm-dependent utility of the proposer is defined by

$$U_p(x, A) = x - \phi_p g(|x-1|) \qquad U_p(x, R) = -\phi_p g(|x-1|).$$

Here (x, A) stands for offer x followed by acceptance and (x, R) stands for offer x followed by rejection. The norm-dependent utility of the responder is given in Table 5.

	Accept	Reject
$x > 1$	$U_r(x, A) = 2 - x - \phi_r$	$U_r(x, R) = 0$
$x \leq 1$	$U_r(x, A) = 2 - x$	$U_r(x, R) = -\phi_r$

Table 5: Responder's utility in the Ultimatum Game.

As in the dictator game the norm-dependent utility is the material payoff minus the disutility from the deviation from the norm. For the proposer we maintain the same assumptions on g and ϕ_p as in the dictator case. For the discrete choice of the responder we assume that he loses utility $\phi_r \geq 0$ if his action is not in accordance with the norm (this is without loss of generality; see Kimbrough et al., 2014).

Now we are ready to characterize the Subgame Perfect Nash Equilibrium (SPNE). When $x \leq 1$, the responder always chooses to accept $(2 - x \geq -\phi_r)$. When $x > 1$ and $\phi_r \leq 1$, the responder will accept all offers $2 - x \geq \phi_r$ and reject all smaller offers. For $\phi_r > 1$, the responder will reject all offers $(2 - x)$ below 1 and accept all offers at or above 1. In other words, the responder accepts an offer $(2 - x)$ if $x \leq x_r^*(\phi_r) = \max\{2 - \phi_r, 1\}$.

The proposer takes into account this best response of the responder. Let

$$x_p^*(\phi_p) = \operatorname*{argmax}_x x - \phi_p g(|x-1|).$$

$x_p^*(\phi_p)$ is in the interval $[1, 2]$. There are two possibilities:

if $\boldsymbol{x_p^*(\phi_p) \leq x_r^*(\phi_r)}$ the proposer chooses $x_p^*(\phi_p)$, i.e. the allocation that maximizes his utility;

if $\boldsymbol{x_p^*(\phi_p) > x_r^*(\phi_r)}$ the proposer chooses $x_r^*(\phi_r)$, or the smallest allocation that will be accepted by the responder.[1]

Note that a similar argument can be used to construct an SPNE under an alternative norm. Consider the "earned rights" treatment from Hoffman et al. (1994), in which the authors observe lower offers and decreased rejection rates when the right to give the ultimatum has been earned. This could be rationalized under an alternative norm induced by the treatment in which it is socially appropriate for proposers to make low offers and for responders to accept a smaller share of an earned pie.

The SPNE just described have an important implicit assumption: all the parameters of the norm-dependent utilities should be common knowledge. In particular, the proposer should exactly know the norm sensitivity ϕ_r of the responder. This sounds rather unrealistic. Let us instead assume that the proposer holds a belief that the responder's ϕ_r is distributed according to some cdf F. Then it can be

[1] Note that our model predicts that selfish behavior by the proposer is increasing in the stakes, because responders will be willing to accept a smaller share of a larger total pie, consistent with the evidence reported in Andersen et al. (2011).

shown that in a Perfect Bayesian Nash Equilibrium the proposer's optimal behavior changes to

$$x^*(\phi_p) = \underset{x\in[0,2]}{\operatorname{argmax}} F(2-x)x - \phi_p g(|x-1|).$$

Here, the best responses of the responder stay the same and do not depend on any incompleteness of information regarding ϕ_r or ϕ_p.

Crucial for our experiment is the implication that the proposers' behavior depends on his beliefs about the responder's type. In some cases, low-ϕ types may want to send large amounts if they believe their counterpart is a high-ϕ type. On the other hand, high-ϕ responders will always be more likely to reject low offers.

Trust Game. In the trust game (Berg et al., 1995) the proposer decides to keep $x \in [0,1]$ and send $1-x$ to the responder. The responder receives $3(1-x)$ and then chooses $y \in [0, 3(1-x)]$, the amount she wants to return to the proposer. The payoff of the proposer is $x+y$ and the payoff of the responder is $3(1-x)-y$. Suppose that the norm is for the proposer to send everything to the responder ($x=0$) and for the responder to send back some amount $y = r_x(1-x)$, where the fraction $r_x \in [0, 1.5]$ is weakly decreasing in x. Here we assume that the responder reciprocates by returning a (weakly) higher fraction of the offer $(1-x)$, the higher the offer is.

The norm-dependent utilities of the proposer and the responder are

$$U_p(x,y) = x + y - \phi_p g(x); \qquad U_r(x,y) = 3(1-x) - y - \phi_r g(\|y - r_x(1-x)\|).$$

Here $\|y - r_x(1-x)\| = |(y - r_x(1-x))/((3-r_x)(1-x))|$ is the normalization that is necessary to keep the deviations from the norm in the $[0,1]$ interval, ensuring that the highest possible disutility from deviation from the norm is equal to ϕ_r in all of the responder's subgames.

In the SPNE the responder chooses

$$y^*(x,\phi_r) = \underset{y\leq 3(1-x)}{\operatorname{argmax}} \; 3(1-x) - y - \phi_r g(\|y - r_x(1-x)\|)$$

which weakly increases in ϕ_r with $y^*(x,\phi_r) \to 0$ as $\phi_r \to 0$ and $y^*(x,\phi_r) \to r_x(1-x)$ as $\phi_r \to \infty$. The proposer takes the best responses of the responder into account and chooses

$$x^*(\phi_p,\phi_r) = \underset{x\in[0,1]}{\operatorname{argmax}} \; x - \phi_p g(x) + y^*(x,\phi_r).$$

Higher values of both ϕ_p and ϕ_r push the optimal proposal towards $x=0$ (i.e. send everything).

Introducing incomplete information into the trust game is not particularly difficult. Here the responder's optimal strategy is unchanged when we introduce uncertainty about ϕ_p. But the proposer now chooses an optimal offer x that solves

$$\max_{x\in[0,1]} \; x - \phi_p g(x) + \int_0^\infty y^*(x,\phi_r) dF(\phi_r)$$

where F is the proposer's belief regarding the responder's norm sensitivity parameter ϕ_r.

Notice that the payoff of the proposer $x - \phi_p g(x) + \int_0^\infty y^*(x,\phi_r) dF(\phi_r)$ might be easily decreasing in x if he believes that ϕ_r is high enough. Therefore, there is no reason to expect strong behavioral differences between high-ϕ and low-ϕ proposers here, since the optimal amount sent depends on beliefs about ϕ_r. On

the other hand, high-ϕ responders will always behave more reciprocally than low-ϕ responders.

Other Games. We note that the general formulation of the model in Kimbrough et al. (2014) also allows us to explain behavior in "distribution games" such as those studied in Engelmann and Strobel (2004), where own payoff is unaffected by own actions; in such a setting, agents with $\phi > 0$ will choose the distribution that accords with the norm, while agents with $\phi = 0$ are indifferent between all distributions. It should also be clear that the model can account for third-party punishment such as that observed in Fehr and Fischbacher (2004), where costly punishment trades off own costs against the disutility of violating a punishment norm.

A.2 Repeated Public Goods Game

Suppose we have an n-player repeated, linear, voluntary contributions public goods game with T periods. $x_{it} \in [0, 1]$ is the action chosen by player i in period t. The material payoff to player i in period t is

$$\pi_i(x_{it}, x_{-it}) = 1 - x_{it} + \alpha \sum_{j=1..n} x_{jt} = 1 - (1 - \alpha)x_{it} + \alpha \sum_{j \neq i} x_{jt}$$

where $\alpha < 1$ and $\alpha n > 1$ (i.e. the payoff from full cooperation is larger than the payoff from full defection). Here the action set $[0, 1]$ has a natural interpretation: 0 is the selfish action, corresponding to keeping the entire endowment in the private account; 1 is the cooperative action. The distance between actions is defined as the distance between corresponding real numbers. The utility of player i in period 1 is

$$u_{i1}(x_{i1}, x_{-i1}) = \pi_i(x_{i1}, x_{-i1}) - \phi_i g(\|\eta - x_{i1}\|).$$

Here ϕ_i and g are the same objects as in the previous section and η represents the action considered the most socially appropriate by all players (in all nodes).[2]

Now we come to an important difference between repeated and one-shot games. We hypothesize that the socially appropriate action in any period depends on the prior play of others. Thus, we introduce a *norm of conditional cooperation* to formalize the intuition that people are not blind adherents to some action, who will continue to choose it under any circumstances. Instead, individuals choose actions *conditional* on past events. Note this is consistent with our definition of a norm, which specifies the appropriate action in all information sets.

We model conditional cooperation in the Public Goods game in a very simple way. We suppose that after each period, each player's appropriate action depends on the average contribution of all players $m_t = \sum_{j \in N} x_{jt}/n$.[3] We alter the utility function to introduce dependence on prior actions. We define the utility of player i in period $t > 1$ as

$$u_{it}(x_{it}, x_{-it}) = \pi_i(x_{it}, x_{-it}) - \phi_i p(m_{t-1}) g(\|\eta - x_{it}\|).$$

We have added a new element to the utility function: the conditional cooperation response $p(m_{t-1})$. This

[2] We assume common knowledge of η. In reality, players might disagree (or be not sure) about what η is before the game is played. Nevertheless, they can learn and update their belief about η by playing the game and observing the contributions. Introducing such learning process would make it possible to incorporate to the model the idea that social norms are changeable and should be learned.

[3] This is consistent with the information available to players in the standard VCM public goods game (and thus with our experimental design) in which players know n and learn the total contribution at the end of each period but do not know each individual's contribution.

is a number in $[0,1]$ which depends on the average actions of all players in the previous period m_{t-1}. In the simplest case, consider $p(m) = 1$ if $m = \eta$ and $p(m) = 0$ otherwise. This is the most basic conditional norm which can be interpreted as follows: if, in the previous period, the average contribution to the public good was consistent with the most socially appropriate action η, then in the current period it is inappropriate to deviate from η. If, however, on average the players contributed less (or more) than η, then the player ceases entirely to care about the costs of deviating from η in the current period. This case is extremely simple, and in reality reactions to norm violations are unlikely to be so stark, but it captures the major intuition we wish to convey - that norms in repeated games account for the history of play.[4]

The overall utility of player i in the repeated game is given by the sum of utilities in all periods: $U_i = \sum_{t=1..T} u_{it}$. This describes an extensive game with observable actions Γ which, unfortunately, does not have a repeated game structure anymore, because payoffs between the repetitions are not independent. Our goal is to characterize Nash Equilibria of this game.

Consider the following strategy s_i^0 of player i:

- In period 1 choose $x_i^* := \text{argmax}_{x\in[0,1]} \ -(1-\alpha)x - \phi_i g(\|\eta - x\|)$;
- After history h contribute
  $x_i^*(h) := \text{argmax}_{x\in[0,1]} \ -(1-\alpha)x - \phi_i p(m(h)) g(\|\eta - x\|)$.

Here $m(h)$ is the average contribution after the last period of history h. This strategy corresponds to each player maximizing her utility in each period separately. Notice that $x_i^* \in [0,\eta]$, since for the values $x \geq \eta$ the function $-(1-\alpha)x - \phi_i g(\|\eta - x\|)$ is strictly decreasing in x. Also, $x_i^*(\phi_i)$ weakly increases in ϕ_i with $x_i^*(\phi_i) = 0$ for $\phi_i \leq \underline{\phi}$ and $x_i^*(\phi_i) = \eta$ for $\phi_i \geq \overline{\phi}$. $x_i^*(\phi_i)$ is strictly increasing on the interval $[\underline{\phi}, \overline{\phi}]$.[5]

Before we present our Propositions let us introduce some notation and a Lemma. For proofs of the Lemma and all Propositions below, see Appendix A.3. Let $\Sigma := \sum_{j\in N} x_j^*$ and $\Sigma_{-i} := \sum_{j\neq i} x_j^*$.

Lemma A.1. Suppose $1/n < \eta$ and consider any subgame of Γ starting in period $t \in 2..T$ such that $m_{t-1} \neq \eta$. Then s^0, restricted to this subgame, is a NE.

In the NE of the subgames described in Lemma A.1 all players choose to contribute zero. It should be noted, however, that, for some values of the parameters, this NE is not unique: another NE exists in which all players contribute η until the penultimate period. We consider such an equilibrium in subgames hardly plausible as it requires full coordination of all players on contributing η *after* some of them have already deviated from it. Nevertheless, the existence of such cooperative NE points to the possibility that deviators *can* return to norm-following given some commonly observed signal, like, for example, a message that focuses attention on cooperative behavior.[6]

Proposition A.1. Suppose $1/n < \eta$ then the following is true

A.1.1 If $\Sigma_{-i} < n\eta - 1$ for all $i \in N$ then the strategy profile $s^0 = (s_i^0)_{i\in N}$ is a NE of Γ.

[4] Note that we assume that contributions greater than prescribed by the norm are also "punished" by abandoning the norm. We assume this to be consistent with the general setup. However, it is easy to alter the $p(\cdot)$ function so that players do not negatively reciprocate when others contribute more than prescribed by the norm, and all results that follow would remain unchanged.

[5] For certain g functions it is possible that $x_i^*(\phi_i)$ does not have either a flat 0 or flat η part (or both).

[6] Engel and Kurschilgen (2013) study public goods games in which they ask questions regarding the beliefs of the subjects before play. They find that asking subjects about the "norm" that should be followed makes them more cooperative, which is in line with our hypothesis.

A.1.2 If $n\eta - 1 \leq (1-\alpha)n\eta$ and $\Sigma_{-i} \leq (1-\alpha)n\eta$ for all $i \in N$ then s^0 is a NE of Γ.

A.1.3 If $n\eta - 1 > (1-\alpha)n\eta$ and $\Sigma_{-i} \geq n\eta - 1$ for some $i \in N$ then s^0 is not a NE of Γ.

The strategy s^0, generates a NE in which players maximize their utility in each period, resulting in a collapse of cooperation after the first period. Notice that conditions of Proposition A.1 roughly say that for this equilibrium to obtain, the ϕ_i parameters for all players should be low enough (this is expressed in terms of x_i^* which are increasing in ϕ_i). The important thing to notice is that this equilibrium *cannot* be sustained if some players have sufficiently high ϕ.

Next we identify two other NE of the game in which cooperation is sustained. Consider the following strategy s_i^1 of player i:

- In period 1 contribute η;
- After history h in period $t < T$ if $m(h) = \eta$ contribute η;
- After history h in period $t < T$ if $m(h) \neq \eta$ contribute $x_i^*(h)$;
- After history h in period T contribute $x_i^*(h)$.

Proposition A.2. Suppose $1/n < \eta$. If for all i it is true that $\Sigma_{-i} \geq \eta\,[1/\alpha - 1]$ then the strategy profile $s^1 = (s_i^1)_{i \in N}$ constitutes a NE of Γ.

Proposition A.2 says that if all players have sufficiently high ϕ, then there is an equilibrium in which, on the equilibrium path, cooperation is sustained at the level of η until the penultimate period.

Next we give an intermediate result for partial cooperation. Consider strategy $s_i^{2\ell}$ of player i:

- In periods 1 to $\ell < T - 1$ contribute η;
- After all histories h not on the path generated by the above rule contribute $x_i^*(h)$.

Proposition A.3. Suppose $1/n < \eta$. If for all i it is true that $\Sigma_{-i} \geq \eta\,[1/\alpha - 1]$ and conditions of Proposition A.1.1 or A.1.2 are satisfied then the strategy profile $s^{2\ell} = (s_i^{2\ell})_{i \in N}$ constitute NE of Γ.

Proposition A.3 indicates that if the parameters of the model are such that both cooperative and non-cooperative equilibria can be sustained, then there is also a collection of other equilibria in which cooperation can be sustained for any fixed number of periods.

In addition, the previous two Propositions indicate that we should expect to observe increasing cooperation as α increases in the interval [0,1). Note that this is consistent with extensive evidence from public goods games despite being inconsistent with the standard payoff maximizing model (e.g. Isaac and Walker, 1988).

Now we briefly describe two Subgame Perfect equilibria that can be constructed using Propositions A.1 and A.2.

Proposition A.4. Suppose that $1/n < \eta$ and the assumptions of Proposition A.1.1 or A.1.2 hold. Then

strategy profile s^0 is a SPNE of Γ.

Proposition A.5. Suppose $1/n < \eta$. If for all i it is true that $\Sigma_{-i} \geq \eta\,[1/\alpha - 1]$ then strategy profile s^1 is a SPNE of Γ.

Finally, we discuss the effect of incomplete information in the repeated public goods game. In Appendix A.4 we show that the cooperative and non-cooperative SPNE just described also constitute Perfect Bayesian Nash Equilibria with essentially no additional assumptions on the uncertainty over ϕ. The reason for this result is that incomplete information does not really impact players' payoff maximizing choices, since they are independent of the characteristics of others, which enter expected utility only indirectly through others' actions.

Result A.1. *Propositions A.1, A.2 and A.3 describe equilibrium behavior in the repeated public goods game:*

- *For low enough norm sensitivity parameters $(\phi_i)_{i\in N}$, which satisfy the inequalities*

$$\sum_{j\neq i} x_j^* < \min\left\{\eta\,[1/\alpha - 1]\,, n\eta - 1\right\}$$

  for all $i \in N$, only the non-cooperative equilibrium exists in which on the equilibrium path players contribute their optimal one-shot game amounts in the first period and then contribute 0 in all other periods;

- *If parameters $(\phi_i)_{i\in N}$ are sufficiently high and satisfy $\sum_{j\neq i} x_j^* > \max\left\{(1-\alpha)n\eta, n\eta - 1\right\}$ for all $i \in N$ only the fully cooperative equilibrium exists in which all players contribute η in all periods but the last, where they maximize the one-shot game payoff;*

- *For the intermediate values of parameters $(\phi_i)_{i\in N}$ cooperative equilibria exist in which cooperation can last from 1 period to $T-1$ periods with later-period defection to 0 as in the non-cooperative equilibrium.*

Thus, the SPNE in which players follow the norm by contributing η only obtains if ϕ_i's and the resulting x_i^* are sufficiently high. In other words, in groups where all members are sufficiently concerned about social norms, there is a SPNE in which the normative action is sustained in the repeated game for all but the final period. Thus, if we can identify norm-sensitive (high-ϕ) and norm-*insensitive* (low-ϕ) types *ex ante* and assortatively match them, we should observe sustained cooperation over time in the high-ϕ groups, even without punishment.

A.3 Public Goods Game Proofs

Proof of Lemma A.1. According to s^0, since $m_{t-1} \neq \eta$, in period t all players should contribute 0 and continue doing so until the end of the game on the s^0-induced path. If some player i deviates in any period, then she can contribute a maximum of 1. Given fixed contributions of others, this will make the average contribution next period equal to $1/n$, which, by assumption, is less than η. Therefore, the deviation will fail to induce norm-following in the next period, and all players will have the same standard stage utility as in the game without norm-dependence. This implies that no single player deviation in any period

(or multiple periods for that matter) can be profitable since the stage utility without norm-dependence decreases in contribution. ■

Proof of Proposition A.1. Notice that from the assumption in A.1.1 it follows that $\Sigma/n < \eta$ (since $x_i^* \leq 1$). The same holds for 1.2: summing up n conditions $\Sigma_{-i} \leq (1-\alpha)n\eta$ for all i gives

$$\Sigma/n \leq \frac{n-\alpha n}{n-1}\eta < \eta$$

where the last inequality follows from the game specific assumption $n\alpha > 1$. This means that in the first period, in accordance with s^0, each player j chooses x_j^* which results in $m_1 = \Sigma/n < \eta$. Therefore, in period 2 all players reciprocate against the norm-violators and stop following the norm which results in 0 contributions by all players in period 2 and thereafter (on the s^0-induced path).

Case A.1.1. Deviation in period 1. By assumption $\Sigma_{-i} < n\eta - 1$ for all $i \in N$. This implies that, if all players $j \neq i$ choose x_j^*, in accordance with s_j^0, then player i cannot choose an x_{i1} sufficiently large that the average contribution reaches η (since $x_{i1} \leq 1$). This means that player i cannot induce norm-following in period 2 and, given any contribution by i, all players will contribute 0 in the next period. Therefore, whatever player i does in period 1, the continuation of the game is the same. Thus, he should follow s_i^0.

Case A.1.2. Deviation in period 1. Suppose that there is player i with $\Sigma_{-i} \geq n\eta - 1$.[7] Here player i can contribute $x > x_i^*$ in order to make the average of contributions Σ/n after period 1 to be equal to η ($x = n\eta - \Sigma_{-i}$).[8] This will encourage norm-following in the next period so that all other players j contribute x_j^* in period 2. Suppose that player i deviates in this way in the first k periods, thus inducing norm-following by the others and then switches back to s_i^0, which maximizes his utility given the choices of others. The payoff on the s^0 path is

$$1 - x_i^* + \alpha\Sigma - \phi_i g(\|\eta - x_i^*\|) + T - 1.$$

The payoff with the described deviation is

$$k[1 - x + \alpha(\Sigma_{-i} + x) - \phi_i g(\|\eta - x\|)] + 1 - x_i^* + \alpha\Sigma - \phi_i g(\|\eta - x_i^*\|) + T - k - 1.$$

No deviation occurs when

$$-x + \alpha(\Sigma_{-i} + x) - \phi_i g(\|\eta - x\|) \leq 0.$$

Substituting x and rearranging we obtain

$$\Sigma_{-i} \leq (1-\alpha)n\eta + \phi_i g(\|\eta - n\eta + \Sigma_{-i}\|).$$

Now, for this to be an equilibrium for all ϕ, including $\phi = 0$, we must have

$$\Sigma_{-i} \leq (1-\alpha)n\eta$$

[7] Otherwise we are back to Case A.1.1.

[8] Deviating to any larger x will only decrease i's payoff without changing anything else. Deviating to less than x will not activate norm-following by others and will leave i with less payoff than he could acquire under s_i^0. Thus, x is the best deviation that activates norm-following.

which is true by the assumption of the Proposition and holds for all $k \leq T$. Since there are only two types of behavior exhibited by other players (contribute x_j^* for all $j \neq i$ or contribute nothing) the described deviation is the best possible for player i. Since this deviation is not profitable, under the assumption, we can conclude that there are no profitable deviations in the first period.

For the Cases A.1.1 and A.1.2 it is left to show that no player i would like to deviate in periods after the first. But this is guaranteed by Lemma A.1.

In Case A.1.3 there is player i with $\Sigma_{-i} \geq n\eta - 1$ who can, as in Case A.1.2, deviate to $n\eta - \Sigma_{-i}$ in period 1. This deviation will be now profitable since in Case A.1.3 $n\eta - 1 > (1-\alpha)n\eta$ and the no deviation condition is $\Sigma_{-i} \leq (1-\alpha)n\eta$. This completes the proof. ■

Proof of Proposition A.2. On the path induced by s^1 all players contribute η in periods 1 to $T-1$ and x_j^* for all $j \in N$ in period T. There can be no deviation in the last period. Therefore, we need to check that no player wants to deviate in periods 1 through $T-1$. The payoff of player i on the s^1-induced path is

$$(T-1)(1-\eta+\alpha n\eta) + 1 - x_i^* + \alpha\Sigma - \phi_i g(\|\eta - x_i^*\|). \tag{1}$$

Now we need to identify the best possible deviation of player i in period k. Any deviation in period k to some contribution less than η will terminate norm-following, and in the next period all other players will contribute 0, according to s_j^0 for all $j \neq i$.[9] Therefore, player i's best choice is to deviate to x_i^*. In all later periods, all players will contribute 0, and player i's best response is to also contribute 0 in all periods after k (by Lemma A.1). Thus, the total payoff received by i from her best deviation to x_i^* in period $k \in 1 \dots T-1$ and later to 0 generates

$$(k-1)(1-\eta+\alpha n\eta) + 1 - x_i^* + \alpha[\eta(n-1) + x_i^*] - \phi_i g(\|\eta - x_i^*\|) + T - k. \tag{2}$$

There is no deviation if

$$\Sigma_{-i} \geq \frac{(T-k)\eta - (T-k-1)\alpha\eta n - \alpha\eta}{\alpha}.$$

Since, by assumption, $\alpha n > 1$ the RHS is increasing in k. So for this inequality to hold for all k we need it to hold for $k = T-1$ which is:

$$\Sigma_{-i} \geq \frac{\eta - \alpha\eta}{\alpha} = \eta\left[\frac{1}{\alpha} - 1\right].$$

The best deviation from s_i^1 is unprofitable if this condition is satisfied. This completes the proof. ■

Proof of Proposition A.3. On the $s^{2\ell}$-induced path all players contribute η in the first ℓ periods, then contribute x_j^* for all $j \in N$ in period $\ell+1$ and contribute 0 thereafter. The proof is analogous to the one of Proposition A.2. The payoff on the $s^{2\ell}$-induced path is

$$\ell(1-\eta+\alpha n\eta) + 1 - x_i^* + \alpha\Sigma - \phi_i g(\|\eta - x_i^*\|) + T - \ell - 1.$$

[9] Deviation to some contribution greater than η is not profitable, as it induces the same behavior in others and decreases i's stage payoff instead of increasing it.

The best deviation during the cooperative stage in period $k \in 1..\ell$ is

$$(k-1)(1-\eta+\alpha n\eta)+1-x_i^*+\alpha[\eta(n-1)+x_i^*]-\phi_i g(\|\eta-x_i^*\|)+T-k.$$

The reasons why this deviation is the best are the same as in Proposition A.2. The deviation is not profitable if

$$\Sigma_{-i} \geq \frac{-\eta(\ell-k)(\alpha n-1)+\eta-\alpha\eta}{\alpha}.$$

The RHS is increasing in k and decreasing in ℓ. Thus the RHS is highest when $k=\ell$. So we obtain the same condition as in Proposition A.2:

$$\Sigma_{-i} \geq \eta\left[\frac{1}{\alpha}-1\right].$$

Now we need to check that there are no profitable deviations in periods after ℓ where the players contribute x_j^* and then 0. By Proposition A.1 we know that such profitable deviations do not exist. This completes the proof. ■

Proof of Proposition A.4. We can classify all non-terminal histories h of Γ by the average contribution $m(h)$. There are two cases: $m(h)=\eta$ and $m(h)\neq\eta$. Strategy profile s^0 is a NE of Γ by Proposition A.1. Strategy profile $s^0(h)$ restricted to subgame $\Gamma(h)$ with $m(h)=\eta$ is the same as s^0 only with less periods. Therefore, by Proposition A.1, $s^0(h)$ is a NE of $\Gamma(h)$. Strategy profile $s^0(h)$ restricted to subgame $\Gamma(h)$ with $m(h)\neq\eta$ is a NE of $\Gamma(h)$ by Lemma A.1. Therefore, s^0 restricted to any subgame is a NE of that subgame, i.e. s^0 is a SPNE of Γ. ■

Proof of Proposition A.5. We can classify all non-terminal histories h of Γ by the average contribution $m(h)$. There are two cases: $m(h)=\eta$ and $m(h)\neq\eta$. Strategy profile s^1 is a NE of Γ by Proposition A.2. Strategy profile $s^1(h)$ restricted to subgame $\Gamma(h)$ with $m(h)=\eta$ is the same as s^1 only with fewer periods. Therefore, by Proposition A.2, $s^1(h)$ is a NE of $\Gamma(h)$. Strategy profile $s^1(h)$ restricted to subgame $\Gamma(h)$ with $m(h)\neq\eta$ is a NE of $\Gamma(h)$ by Lemma A.1. Therefore, s^1 restricted to any subgame is NE or, i.e. s^1 is a SPNE of Γ. ■

A.4 Repeated Public Goods Game with Incomplete Information

In this section we introduce incomplete information into the game Γ. The reason for doing this is that the norm sensitivity parameters ϕ, are not only unobservable but also may be very hard to deduce from observed behavior. For example, in the setup of Proposition A.2 players with a wide range of ϕ's choose to contribute η in equilibrium, which would not allow observers to estimate their ϕ.

We represent incomplete information as uncertainty about x_i^*, the optimal choice of player i in one-shot game.[10] Assume that, for each player i, x_i^* is distributed according to some F_i. All F_i are common knowledge. In what follows we will construct two Perfect Bayesian Nash Equilibria (PBNE) of Γ with incomplete information structure given by $(F_i)_{i\in N}$ which correspond to the SPNE depicted in Propositions A.4 and A.5.

[10] We could have defined uncertainty directly over ϕ_i, but we find it more convenient to work with uncertainty over x_i^*. The two formulations are equivalent: ϕ_i's that correspond to some x_i^* can be found through the inverse of $x_i^*(\phi_i)$.

It is easier to start with the cooperative equilibrium (Proposition A.5). The strategy profile s^1 will also constitute a PBNE. We only need to introduce common beliefs about others' types to each history. For histories h with $m(h) = \eta$ let $\mu_i(h) = F_i$ be the common belief of all players but i about her type after h. For each history with $m(h) \neq \eta$ let $\mu_i(h) = \delta_{x_i(h)}$ be the point mass at $x_i(h)$, the choice of i in the last period of h. Thus, we assume that whenever a player observes someone choosing some contribution $x < \eta$ she believes that this player has chosen according to her one-shot game best reply. In principle, the beliefs can be defined in many other ways, but for our particular game this is irrelevant. The only thing that matters is that the prior common belief is $(F_i)_{i\in N}$ and that this belief does not get updated if all players choose η (see Osborne and Rubinstein (1994) Section 12.3 for the relevant definitions).

Before we proceed let us give some definitions. Let $E := \sum_{j\in N} E_{F_j}[x_j^*]$ be the sum of expected values of x_j^* given F_j and let E_{-i} be the same sum without player i.

Proposition A.6. Suppose $1/n < \eta$. If $E_{-i} \geq \eta\,[1/\alpha - 1]$ for all i then the strategy profile s^1 together with the beliefs $(\mu_i(h))_{i\in N, h\in H\setminus Z}$ constitute a PBNE.

Proof. The proof that no type of any player wants to deviate from the strategy s^1 in Γ is exactly the same as that of Proposition A.2. Only Σ_{-i} is replaced with E_{-i} and Σ is replaced with $E_{-i} + x_i^*$. Similarly, no player wants to deviate from $s^1(h)$ in subgames h with $m(h) \neq \eta$ by Lemma A.1: uncertainty plays no role here. In all subgames $s^1(h)$ with $m(h) = \eta$ the proof of Proposition A.2 is used again since by assumption the beliefs after any such history are $(F_i)_{i\in N}$ as in the supergame. ■

Proposition A.7. Suppose that $1/n < \eta$ and the assumptions of Proposition A.1.1 or A.1.2 hold with Σ_{-i} replaced with E_{-i}. Then the strategy profile s^0 together with the beliefs $(\mu_i(h))_{i\in N, h\in H\setminus Z}$ constitute a PBNE.

Proof. The proof that no type of any player wants to deviate from the strategy s^0 in Γ is exactly the same as that of Proposition A.1. Only Σ_{-i} is replaced with E_{-i} as above and Σ is replaced with $E_{-i} + x_i^*$. Similarly, no player wants to deviate from $s^0(h)$ in subgames h with $m(h) \neq \eta$ by Lemma A.1: uncertainty plays no role here. In all subgames $s^0(h)$ with $m(h) = \eta$ the proof of Proposition A.2 is used again since by assumption the beliefs after any such history are $(F_i)_{i\in N}$ as in the supergame. ■

B Instructions

B.1 Instructions for the Rule Following Stage

General information
You are now participating in a decision making experiment. If you follow the instructions carefully, you can earn a considerable amount of money depending on your decisions and the decisions of the other participants. Your earnings will be paid to you in CASH at the end of the experiment

This set of instructions is for your private use only. **During the experiment you are not allowed to communicate with anybody**. In case of questions, please raise your hand. Then we will come to your seat and answer your questions. Any violation of this rule excludes you immediately from the experiment and all payments. The research organization METEOR has provided funds for conducting this experiment.

Part I
In Part I of this experiment, you control a stick figure that will walk across the screen.

Once the experiment begins, you can start walking by clicking the **"Start"** button on the left of the screen. Your stick figure will approach a series of stop lights and will stop to wait at each light. To make your stick figure walk again, click the **"Walk"** button in the middle of the screen.

The rule is to wait at each stop light until it turns green.

Your earnings in Part I are determined by the amount of time it takes your stick figure to walk across the screen. **Specifically, you begin with an initial endowment of 8 Euro**. Each second, this endowment will decrease by **0.08 Euro**.

This is the end of the instructions for Part I. If you have any questions, please raise your hand and an experimenter will answer them privately. Otherwise, please wait quietly for the experiment to begin.

B.2 Instructions for the Public Goods Game

Part II

This part of the experiment will consist of several decision making periods. In each period, you are given an endowment of 50 tokens. Your task is to decide how to divide these tokens into either or both of two accounts: a private account and a group account. Each period you receive the sum of your earnings from your private account plus your earnings from the group account.

There are 4 people, including yourself, participating in your group. You will be matched with the same people for all of Part II.

Each token you place in the private account generates a cash return to you (and to you alone) of one cent (0.01 Euro). Tokens placed in the group account yield a different return.

Every member of the group receives the same return for each token you place in the group account. Similarly, you receive a return for every token that the other members of the group place in the group account. Thus, your earnings in each decision period are the number of tokens you place in your private account, plus the return from all tokens you and the other members of the group place in the group account.

Specifically, the total amount of tokens in the group account, that is, your group account tokens and the tokens placed in the group account by other members of the group, is doubled and then equally divided among 4 members of the group.

Here are two examples to make this clear:

(1) Suppose you place 0 tokens in the group account and the other members of your group place a total of 150 tokens in the group account. Your earnings from the group account would be (2 * 150) / 4 = 75 cents. Other members of the group would also receive 75 cents from the group account.

(2) Suppose you place 45 tokens in the group account and the other members of your group place a total of 15 tokens in the group account. The total group contribution is 60. Your earnings from the group account would be (2 * 60) / 4 = 30 cents. Other members of the group would also receive 30 cents from the group account.

Each period proceeds as follows:

First, decide on the number of tokens to place in the private and in the group accounts by entering numbers into the boxes labeled private and group. Your entries must sum to your token endowment which is always 50. While you make your decision, the 3 other members in your group will also divide their token endowments between the private and group accounts.

Second, after everyone has made a decision, your earnings for that decision period are the sum of your earnings from the private and group accounts.

As an example, suppose the total contribution to the group account at the end of the period was 120. Your

contribution to the group account was 30, which means your contribution to the private account was 20. You would earn 80 cents this period, 20 from private account and (2 * 120) / 4 = 60 from the group account.

While you are deciding how to allocate your tokens, everyone else in your group will be doing so as well. When the period is over the computer will display your earnings for that period and your total earnings up to and including that period.

This is the end of the instructions. If you have any questions please raise your hand and an experimenter will come by to answer them.

B.3 Instructions for the Trust Game

Part II

This part of the experiment will consist of several periods.

In this part, there will be two types of people, Red and Blue. You will be both a Red person and a Blue person depending on the period. Each period you will be randomly paired with a person of the other type. In this experiment you will interact with 3 other people in the room.

Instructions for Blue People

Each Blue person begins each period with 80 tokens. A Blue person may choose to send some, all, or none of these tokens to a Red person he/she is paired with by typing the amount into a box in the center of the screen and then clicking "OK".

Any tokens that a Blue person sends to a Red person will be subtracted from the Blue persons account, multiplied by 3 and transferred to the Red person. Any tokens that a Blue person chooses not to send to the Red person remain the Blue persons earnings. (Only Blue people will be able to send tokens and have them multiplied.)

Instructions for Red People

Each Red person enters a period with 80 tokens. After the Blue person makes a decision, the Red person will see how many tokens were sent from the Blue person.

The amount sent by the Blue person will be multiplied by 3 and added to the Red persons account. Then the Red person decides to send some, all or none of these tokens to the Blue person by typing the amount into a box in the center of the screen and then clicking "OK". (Only Red people will make this decision.)

In each period, each Red person is paired with one Blue person for the entire period. (One "period" consists of one Blue person deciding how many tokens to send to one Red person and that Red person deciding how many of the multiplied tokens to send to the paired Blue person.)

Summary

A Blue persons earnings for a period are:
Earnings =
Starting tokens
minus Amount Sent to Red
plus Amount Received from Red

A Red persons earnings for a period are:
Earnings =
Starting tokens
plus Amount Received from Blue x 3
minus Amount Sent to Blue

At the end of the experiment the sum of your tokens from all periods will be converted to Euros at a rate of 100 tokens = 1 Euro and paid to you privately in cash, along with your earnings from Part 1 of the experiment. This is the end of the instructions. If you have any questions please raise your hand and an experimenter will come by to answer them.

B.4 Instructions for the Dictator Game

Part II

In this part, there will be two types of people, Red and Blue. Throughout Part II, you will be either a Red person or a Blue person depending on a random choice by the computer.

You will be paired with a person of the other type.

You will interact with only 1 other person in the room. In this experiment a Blue person makes a choice and a Red person does not choose. The amount of money you receive depends on the decision made by the Blue person in your pair.

Instructions for Blue People

Each Blue person begins with 16 Euro. A Blue person chooses how to allocate this money between him/herself and a Red person he/she is paired with. To specify an allocation, the Blue person types the amount he/she wants to allocate to him/herself and the amount he/she wants to allocate to the Red person and then clicks OK. The two amounts must sum up to 16 Euro.

Instructions for Red People

After the Blue person chooses an allocation, the Red person will see how much money the Blue person allocated to the Red person. The Red person does not make any decisions.

This is the end of the instructions for Part II. If you have any questions please raise your hand and an experimenter will come by to answer them.

B.5 Instructions for the Ultimatum Game

Part II

In this part, there will be two types of people, Red and Blue. Throughout Part II, you will be either a Red person or a Blue person depending on a random choice by the computer.

You will be paired with a person of the other type. You will interact with only 1 other person in the room. The amount of money you receive depends on the decision you make and on the decision of the person you are paired with.

Each person makes his/her decision without knowing the decision of the other person.

Instructions for Blue People

Each Blue person begins with 16 Euro. A Blue person chooses how to allocate this money between him/herself and a Red person he/she is paired with. To specify an allocation, the Blue person types the amount he/she wants to allocate to him/herself and the amount he/she wants to allocate to the Red person and then clicks OK. The two amounts must sum up to 16 Euro.

Instructions for Red People

The Red person makes a decision that determines whether he/she will Accept or Reject the allocation chosen by the Blue person. If the Red person accepts an allocation, then he/she gets the amount of money specified by this allocation and the Blue person gets his/her part. If the Red person rejects an allocation, then both the Red and the Blue person get nothing.

In practice, the Red person chooses a single number (call it X). This number represents the minimum amount that the Red person is willing to accept. Once the number has been entered, the Red person will click "OK." Then, if the Blue person chose any amount (call it Y) that is greater than or equal to the Red person's chosen minimum (in other words, if $Y \geq X$), the Red person accepts the allocation and receives Y Euro while the Blue person receives 16 - Y Euro. On the other hand, if Y is less than the Red person's chosen minimum (that is, if $Y < X$), then the allocation is rejected and both persons receive nothing.

After both the Blue and the Red persons have decided, the outcome will be revealed and both people will have their earnings added to their total for the experiment.

This is the end of the instructions for Part II. If you have any questions please raise your hand and an experimenter will come by to answer them.

C Moral Foundations Questionnaire

Part 1. When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please rate each statement using this scale:

0	1	2	3	4	5
not at all relevant	not very relevant	slightly relevant	somewhat relevant	very relevant	extremely relevant

1. Whether or not someone suffered emotionally
2. Whether or not some people were treated differently than others
3. Whether or not someone's action showed love for his or her country
4. Whether or not someone showed a lack of respect for authority
5. Whether or not someone violated standards of purity and decency
6. Whether or not someone was good at math
7. Whether or not someone cared for someone weak or vulnerable
8. Whether or not someone acted unfairly
9. Whether or not someone did something to betray his or her group
10. Whether or not someone conformed to the traditions of society
11. Whether or not someone did something disgusting
12. Whether or not someone was cruel
13. Whether or not someone was denied his or her rights
14. Whether or not someone showed a lack of loyalty
15. Whether or not an action caused chaos or disorder
16. Whether or not someone acted in a way that God would approve of

Part 2. Please read the following sentences and indicate your agreement or disagreement:

0	1	2	3	4	5
Strongly Disagree	Moderately Disagree	Slightly Disagree	Slightly Agree	Moderately Agree	Strongly Agree

1. Compassion for those who are suffering is the most crucial virtue.
2. When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.
3. I am proud of my countrys history.
4. Respect for authority is something all children need to learn.
5. People should not do things that are disgusting, even if no one is harmed.
6. It is better to do good than to do bad.
7. One of the worst things a person could do is hurt a defenseless animal.
8. Justice is the most important requirement for a society.
9. People should be loyal to their family members, even when they have done something wrong.
10. Men and women each have different roles to play in society.
11. I would call some acts wrong on the grounds that they are unnatural.
12. It can never be right to kill a human being.
13. I think its morally wrong that rich children inherit a lot of money while poor children inherit nothing.
14. It is more important to be a team player than to express oneself.
15. If I were a soldier and disagreed with my commanding officers orders, I would obey anyway because that is my duty.
16. Chastity is an important and valuable virtue.

*The Moral Foundations Questionnaire (full version, July 2008) by Jesse Graham, Jonathan Haidt, and Brian Nosek. For more information about Moral Foundations Theory and scoring this form, see: www.MoralFoundations.org

Moral Foundations Questionnaire: 30-Item Full Version Item Key, July 2008

–Below are the items that compose the MFQ30. Variable names are IN CAPS
–Besides the 30 test items there are 2 catch items, MATH and GOOD
–For more information about the theory, or to print out a version of this scale formatted for participants, or to learn about scoring this scale, please see: www.moralfoundations.org

---

PART 1 ITEMS (responded to using the following response options: not at all relevant, not very relevant, slightly relevant, somewhat relevant, very relevant, extremely relevant)

MATH - Whether or not someone was good at math [This item is not scored; it is included both to force people to use the bottom end of the scale, and to catch and cut participants who respond with last 3 response options]

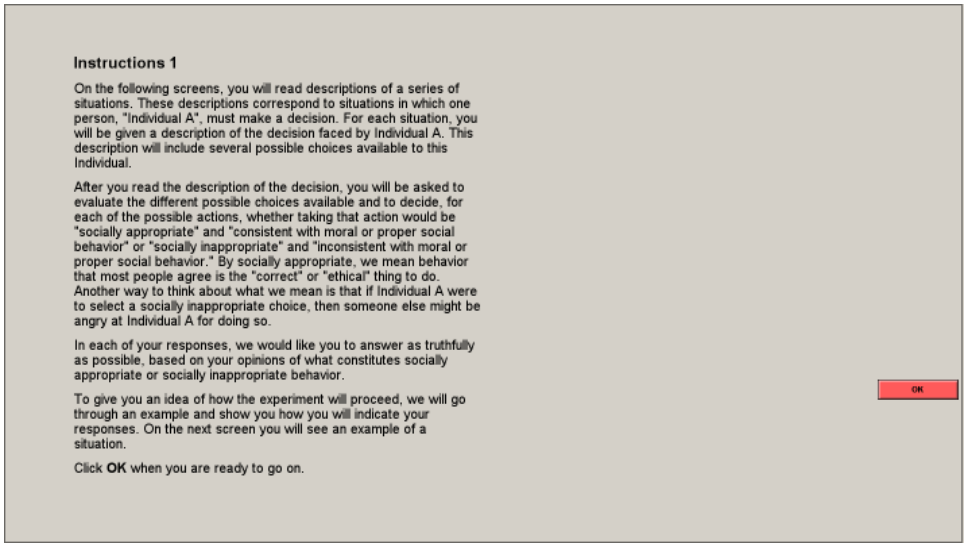
Harm:
EMOTIONALLY - Whether or not someone suffered emotionally
WEAK - Whether or not someone cared for someone weak or vulnerable
CRUEL - Whether or not someone was cruel
Fairness:
TREATED - Whether or not some people were treated differently than others
UNFAIRLY - Whether or not someone acted unfairly
RIGHTS - Whether or not someone was denied his or her rights
Ingroup:
LOVECOUNTRY - Whether someones action showed love for his or her country
BETRAY - Whether or not someone did something to betray his or her group
LOYALTY - Whether or not someone showed a lack of loyalty
Authority:
RESPECT - Whether or not someone showed a lack of respect for authority
TRADITIONS - Whether or not someone conformed to the traditions of society
CHAOS - Whether or not an action caused chaos or disorder
Purity:
DECENCY - Whether or not someone violated standards of purity and decency
DISGUSTING - Whether or not someone did something disgusting
GOD - Whether or not someone acted in a way that God would approve of

PART 2 ITEMS (responded to using the following response options: strongly disagree, moderately disagree, slightly disagree, slightly agree, moderately agree, strongly agree)

GOOD It is better to do good than to do bad. [Not scored, included to force use of top of the scale, and to catch and cut people who respond with first 3 response options]

Harm:
COMPASSION - Compassion for those who are suffering is the most crucial virtue.
ANIMAL - One of the worst things a person could do is hurt a defenseless animal.
KILL - It can never be right to kill a human being.
Fairness:
FAIRLY - When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.
JUSTICE Justice is the most important requirement for a society.
RICH - I think its morally wrong that rich children inherit a lot of money while poor children inherit nothing.
Ingroup:
HISTORY - I am proud of my countrys history.
FAMILY - People should be loyal to their family members, even when they have done something wrong.
TEAM - It is more important to be a team player than to express oneself.
Authority:
KIDRESPECT - Respect for authority is something all children need to learn.
SEXROLES - Men and women each have different roles to play in society.
SOLDIER - If I were a soldier and disagreed with my commanding officers orders, I would obey anyway because that is my duty.
Purity:
HARMLESSDG - People should not do things that are disgusting, even if no one is harmed.
UNNATURAL - I would call some acts wrong on the grounds that they are unnatural.
CHASTITY - Chastity is an important and valuable virtue.

D Instructions for the Norm Elicitation Treatments



Instructions 1

On the following screens, you will read descriptions of a series of situations. These descriptions correspond to situations in which one person, "Individual A", must make a decision. For each situation, you will be given a description of the decision faced by Individual A. This description will include several possible choices available to this Individual.

After you read the description of the decision, you will be asked to evaluate the different possible choices available and to decide, for each of the possible actions, whether taking that action would be "socially appropriate" and "consistent with moral or proper social behavior" or "socially inappropriate" and "inconsistent with moral or proper social behavior." By socially appropriate, we mean behavior that most people agree is the "correct" or "ethical" thing to do. Another way to think about what we mean is that if Individual A were to select a socially inappropriate choice, then someone else might be angry at Individual A for doing so.

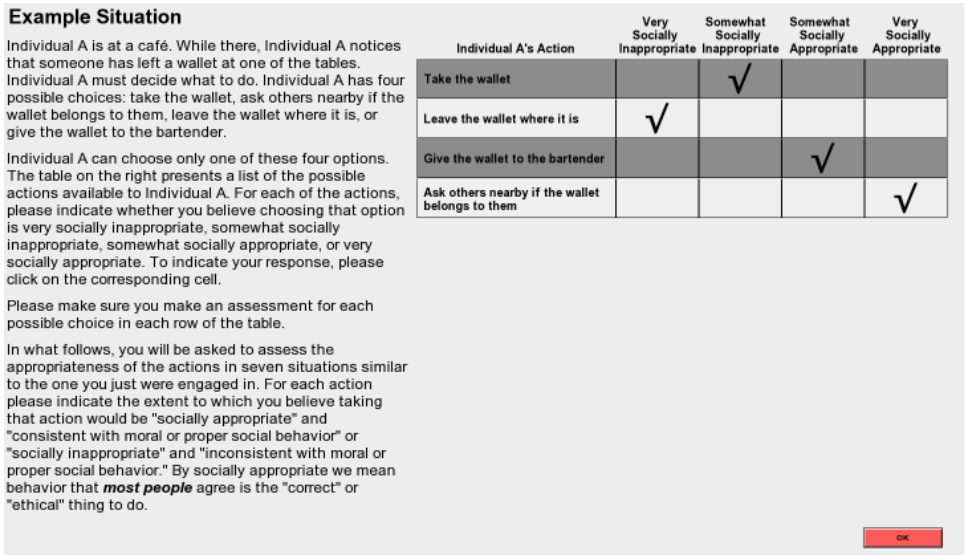
In each of your responses, we would like you to answer as truthfully as possible, based on your opinions of what constitutes socially appropriate or socially inappropriate behavior.

To give you an idea of how the experiment will proceed, we will go through an example and show you how you will indicate your responses. On the next screen you will see an example of a situation.

Click **OK** when you are ready to go on.

OK

Figure D1: All treatments, Page 1.



Example Situation

Individual A is at a café. While there, Individual A notices that someone has left a wallet at one of the tables. Individual A must decide what to do. Individual A has four possible choices: take the wallet, ask others nearby if the wallet belongs to them, leave the wallet where it is, or give the wallet to the bartender.

Individual A can choose only one of these four options. The table on the right presents a list of the possible actions available to Individual A. For each of the actions, please indicate whether you believe choosing that option is very socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, or very socially appropriate. To indicate your response, please click on the corresponding cell.

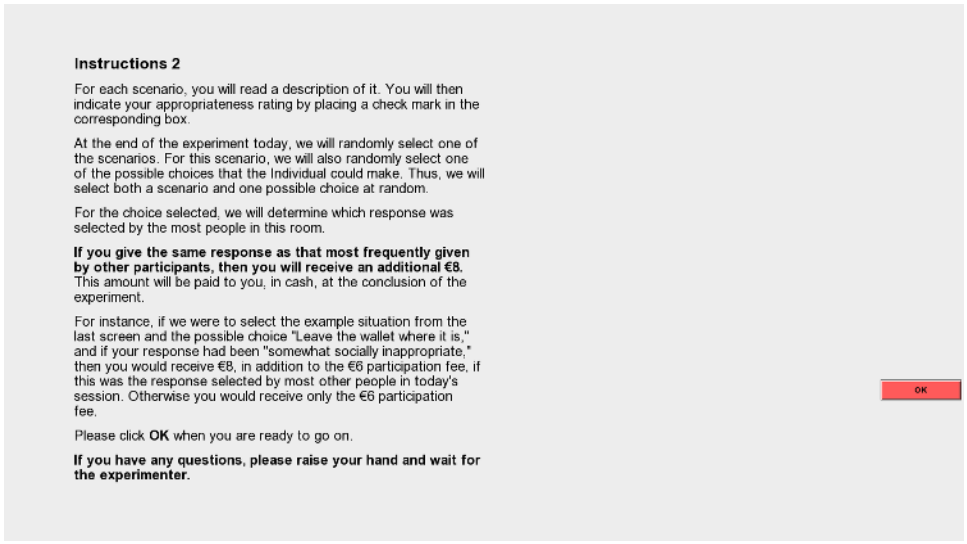
Please make sure you make an assessment for each possible choice in each row of the table.

In what follows, you will be asked to assess the appropriateness of the actions in seven situations similar to the one you just were engaged in. For each action please indicate the extent to which you believe taking that action would be "socially appropriate" and "consistent with moral or proper social behavior" or "socially inappropriate" and "inconsistent with moral or proper social behavior." By socially appropriate we mean behavior that ***most people*** agree is the "correct" or "ethical" thing to do.

Individual A's Action	Very Socially Inappropriate	Somewhat Socially Inappropriate	Somewhat Socially Appropriate	Very Socially Appropriate
Take the wallet		√		
Leave the wallet where it is	√			
Give the wallet to the bartender			√	
Ask others nearby if the wallet belongs to them				√

OK

Figure D2: All treatments, Page 2.



Instructions 2

For each scenario, you will read a description of it. You will then indicate your appropriateness rating by placing a check mark in the corresponding box.

At the end of the experiment today, we will randomly select one of the scenarios. For this scenario, we will also randomly select one of the possible choices that the Individual could make. Thus, we will select both a scenario and one possible choice at random.

For the choice selected, we will determine which response was selected by the most people in this room.

If you give the same response as that most frequently given by other participants, then you will receive an additional €8. This amount will be paid to you, in cash, at the conclusion of the experiment.

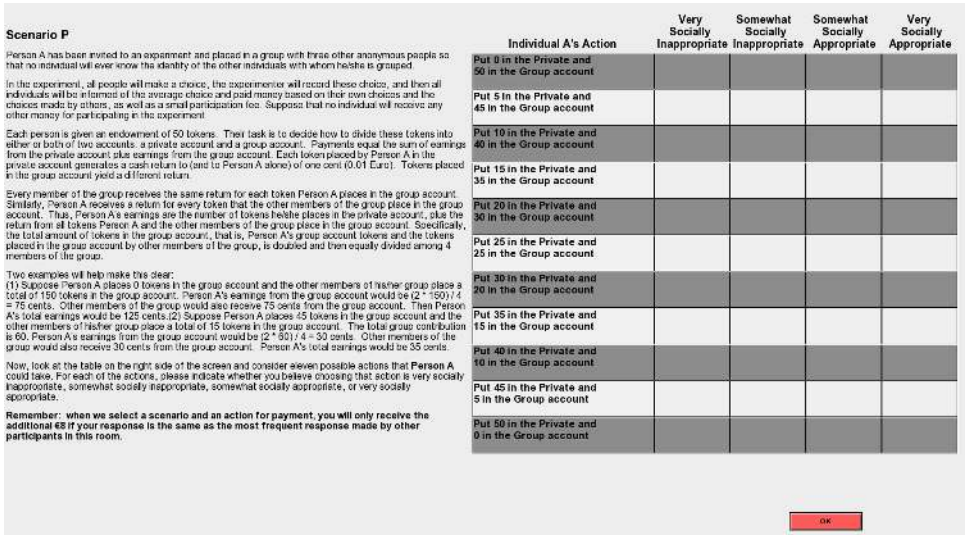
For instance, if we were to select the example situation from the last screen and the possible choice "Leave the wallet where it is," and if your response had been "somewhat socially inappropriate," then you would receive €8, in addition to the €6 participation fee, if this was the response selected by most other people in today's session. Otherwise you would receive only the €6 participation fee.

Please click **OK** when you are ready to go on.

If you have any questions, please raise your hand and wait for the experimenter.

OK

Figure D3: All treatments, Page 3.



Scenario P

Person A has been invited to an experiment and placed in a group with three other anonymous people so that no individual will ever know the identity of the other individuals with whom he/she is grouped.

In the experiment, all people will make a choice, the experimenter will record these choice, and then all individuals will be informed of the average choice and paid money based on their own choices and the choices made by others, as well as a small participation fee. Suppose that no individual will receive any other money for participating in the experiment.

Each person is given an endowment of 50 tokens. Their task is to decide how to divide these tokens into either or both of two accounts: a private account and a group account. Payments equal the sum of earnings from the private account plus earnings from the group account. Each token placed by Person A in the private account generates a cash return to (and to Person A alone) of one cent (0.01 Euro). Tokens placed in the group account yield a different return.

Every member of the group receives the same return for each token Person A places in the group account. Similarly, Person A receives a return for every token that the other members of the group place in the group account. Thus, Person A's earnings are the number of tokens he/she places in the private account, plus the return from all tokens Person A and the other members of the group place in the group account. Specifically, the total amount of tokens in the group account, that is, Person A's group account tokens and the tokens placed in the group account by other members of the group, is doubled and then equally divided among 4 members of the group.

Two examples will help make this clear:
(1) Suppose Person A places 0 tokens in the group account and the other members of his/her group place a total of 150 tokens in the group account. Person A's earnings from the group account would be (2 * 150) / 4 = 75 cents. Other members of the group would also receive 75 cents from the group account. Then Person A's total earnings would be 125 cents.(2) Suppose Person A places 45 tokens in the group account and the other members of his/her group place a total of 15 tokens in the group account. The total group contribution is 60. Person A's earnings from the group account would be (2 * 60) / 4 = 30 cents. Other members of the group would also receive 30 cents from the group account. Person A's total earnings would be 55 cents.

Now, look at the table on the right side of the screen and consider eleven possible actions that **Person A** could take. For each of the actions, please indicate whether you believe choosing that action is very socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, or very socially appropriate.

Remember: when we select a scenario and an action for payment, you will only receive the additional €8 if your response is the same as the most frequent response made by other participants in this room.

Individual A's Action	Very Socially Inappropriate	Somewhat Socially Inappropriate	Somewhat Socially Appropriate	Very Socially Appropriate
Put 0 in the Private and 50 in the Group account				
Put 5 in the Private and 45 in the Group account				
Put 10 in the Private and 40 in the Group account				
Put 15 in the Private and 35 in the Group account				
Put 20 in the Private and 30 in the Group account				
Put 25 in the Private and 25 in the Group account				
Put 30 in the Private and 20 in the Group account				
Put 35 in the Private and 15 in the Group account				
Put 40 in the Private and 10 in the Group account				
Put 45 in the Private and 5 in the Group account				
Put 50 in the Private and 0 in the Group account				

OK

Figure D4: One-Shot PG/DG Elicitation Treatment, Page 4 (for half of the subjects the PG game scenario followed after the Dictator game scenario, see next figure).

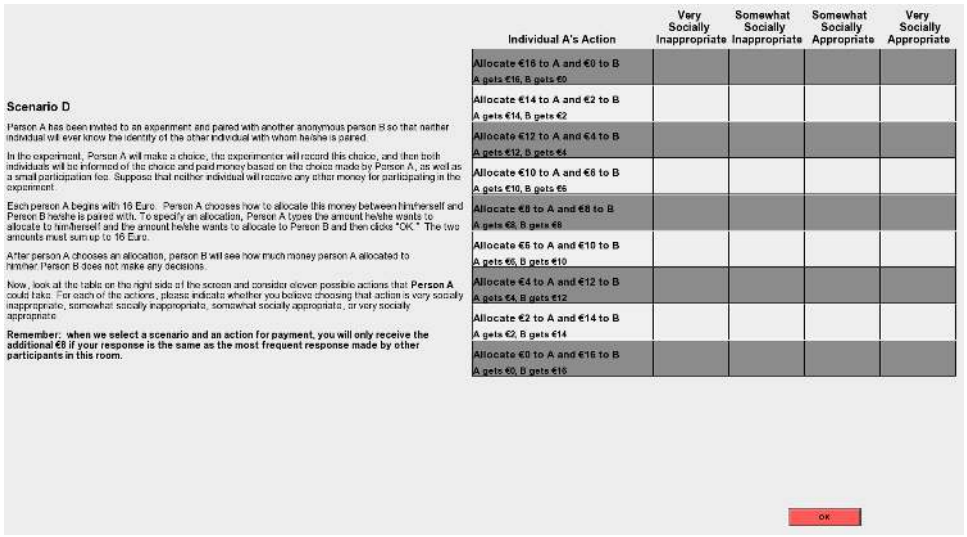


Figure D5: One-Shot PG/DG Elicitation Treatment, Page 5 (for half of the subjects the Dictator game scenario followed before the PG game scenario, see previous figure).

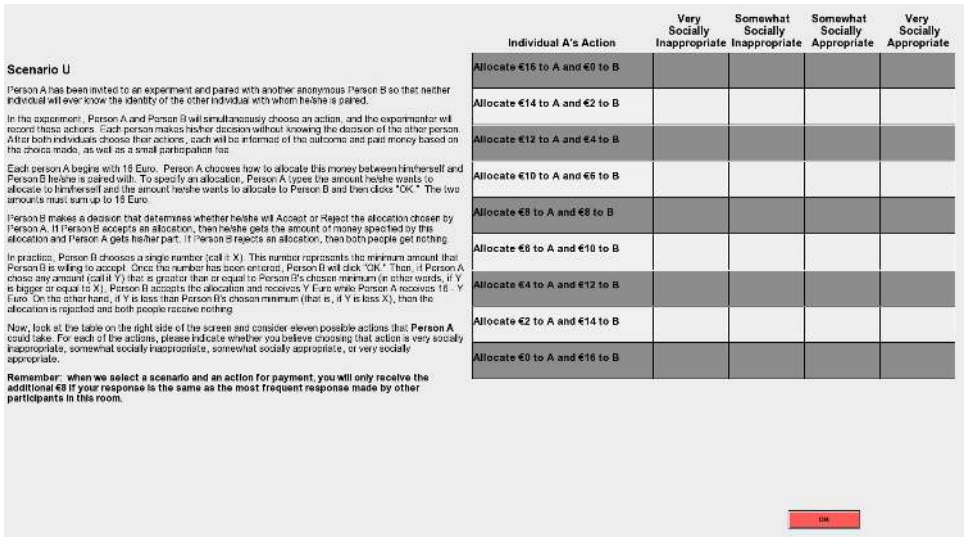
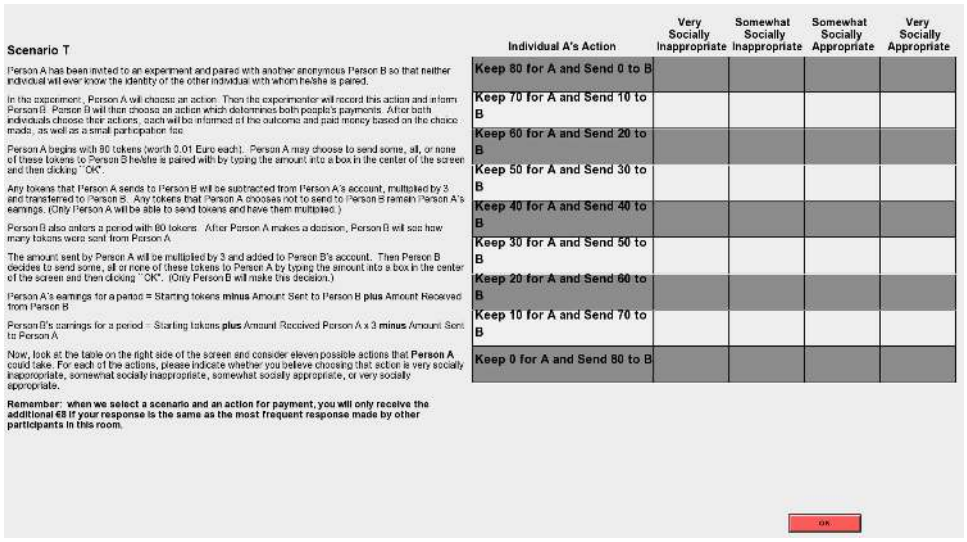


Figure D6: Proposer TG/UG Elicitation Treatment, Page 4 (for half of the subjects the Ultimatum game scenario followed after the Trust game scenarios, see next figure).



Scenario T

Person A has been invited to an experiment and paired with another anonymous Person B so that neither individual will ever know the identity of the other individual with whom he/she is paired.

In the experiment, Person A will choose an action. Then the experimenter will record this action and inform Person B. Person B will then choose an action which determines both people's payments. After both individuals choose their actions, each will be informed of the outcome and paid money based on the choice made, as well as a small participation fee.

Person A begins with 80 tokens (worth 0.01 Euro each). Person A may choose to send some, all, or none of these tokens to Person B he/she is paired with by typing the amount into a box in the center of the screen and then clicking "OK".

Any tokens that Person A sends to Person B will be subtracted from Person A's account, multiplied by 3 and transferred to Person B. Any tokens that Person A chooses not to send to Person B remain Person A's earnings. (Only Person A will be able to send tokens and have them multiplied.)

Person B also enters a period with 80 tokens. After Person A makes a decision, Person B will see how many tokens were sent from Person A.

The amount sent by Person A will be multiplied by 3 and added to Person B's account. Then Person B decides to send some, all or none of these tokens to Person A by typing the amount into a box in the center of the screen and then clicking "OK". (Only Person B will make this decision.)

Person A's earnings for a period = Starting tokens **minus** Amount Sent to Person B **plus** Amount Received from Person B

Person B's earnings for a period = Starting tokens **plus** Amount Received Person A x 3 **minus** Amount Sent to Person A

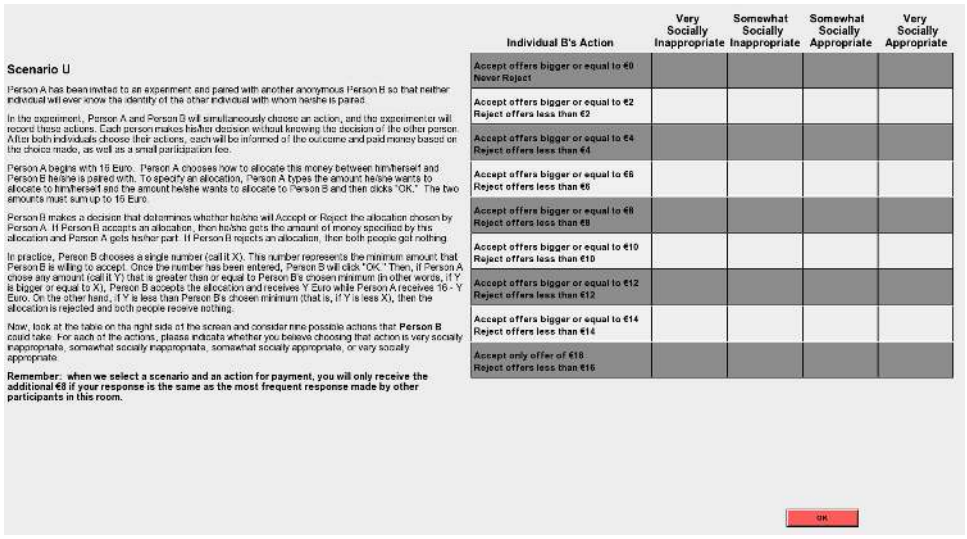
Now, look at the table on the right side of the screen and consider eleven possible actions that **Person A** could take. For each of the actions, please indicate whether you believe choosing that action is very socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, or very socially appropriate.

Remember: when we select a scenario and an action for payment, you will only receive the additional €8 if your response is the same as the most frequent response made by other participants in this room.

Individual A's Action	Very Socially Inappropriate	Somewhat Socially Inappropriate	Somewhat Socially Appropriate	Very Socially Appropriate
Keep 80 for A and Send 0 to B				
Keep 70 for A and Send 10 to B				
Keep 60 for A and Send 20 to B				
Keep 50 for A and Send 30 to B				
Keep 40 for A and Send 40 to B				
Keep 30 for A and Send 50 to B				
Keep 20 for A and Send 60 to B				
Keep 10 for A and Send 70 to B				
Keep 0 for A and Send 80 to B				

OK

Figure D7: Proposer TG/UG Elicitation Treatment, Page 5 (for half of the subjects the Trust game scenarios followed before the Ultimatum game scenario, see previous figure).



Scenario U

Person A has been invited to an experiment and paired with another anonymous Person B so that neither individual will ever know the identity of the other individual with whom he/she is paired.

In the experiment, Person A and Person B will simultaneously choose an action, and the experimenter will record these actions. Each person makes his/her decision without knowing the decision of the other person. After both individuals choose their actions, each will be informed of the outcome and paid money based on the choice made, as well as a small participation fee.

Person A begins with 15 Euro. Person A chooses how to allocate this money between him/herself and Person B he/she is paired with. To specify an allocation, Person A types the amount he/she wants to allocate to him/herself and the amount he/she wants to allocate to Person B and then clicks "OK." The two amounts must sum up to 15 Euro.

Person B makes a decision that determines whether he/she will Accept or Reject the allocation chosen by Person A. If Person B accepts an allocation, then he/she gets the amount of money specified by this allocation and Person A gets his/her part. If Person B rejects an allocation, then both people get nothing.

In practice, Person B chooses a single number (call it X). This number represents the minimum amount that Person B is willing to accept. Once the number has been entered, Person B will click "OK." Then, if Person A chose any amount (call it Y) that is greater than or equal to Person B's chosen minimum (in other words, if Y is bigger or equal to X), Person B accepts the allocation and receives Y Euro while Person A receives 15 - Y Euro. On the other hand, if Y is less than Person B's chosen minimum (that is, if Y is less X), then the allocation is rejected and both people receive nothing.

Now, look at the table on the right side of the screen and consider nine possible actions that **Person B** could take. For each of the actions, please indicate whether you believe choosing that action is very socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, or very socially appropriate.

Remember: when we select a scenario and an action for payment, you will only receive the additional €8 if your response is the same as the most frequent response made by other participants in this room.

Individual B's Action	Very Socially Inappropriate	Somewhat Socially Inappropriate	Somewhat Socially Appropriate	Very Socially Appropriate
Accept offers bigger or equal to €0 Never Reject				
Accept offers bigger or equal to €2 Reject offers less than €2				
Accept offers bigger or equal to €4 Reject offers less than €4				
Accept offers bigger or equal to €6 Reject offers less than €6				
Accept offers bigger or equal to €8 Reject offers less than €8				
Accept offers bigger or equal to €10 Reject offers less than €10				
Accept offers bigger or equal to €12 Reject offers less than €12				
Accept offers bigger or equal to €14 Reject offers less than €14				
Accept only offer of €16 Reject offers less than €16				

OK

Figure D8: Responder TG/UG Elicitation Treatment, Page 4 (for half of the subjects the Ultimatum game scenario followed after the Trust game scenarios, see next figure).

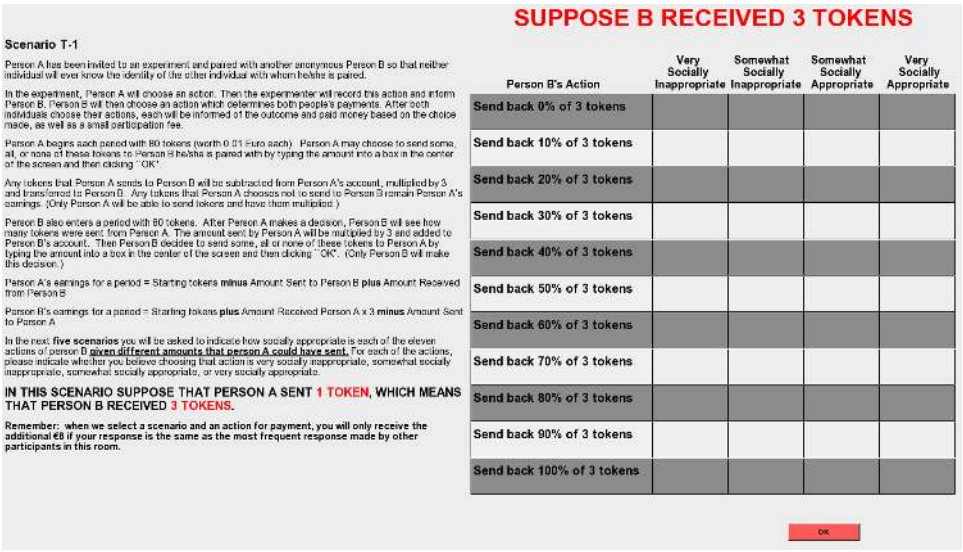


Figure D9: Responder TG/UG Elicitation Treatment, Pages 5-9. The amount of tokens that B received varied on the five pages: 3, 60, 120, 180, 240 tokens (for half of the subjects the Trust game scenarios followed before the Ultimatum game scenario, see previous figure).

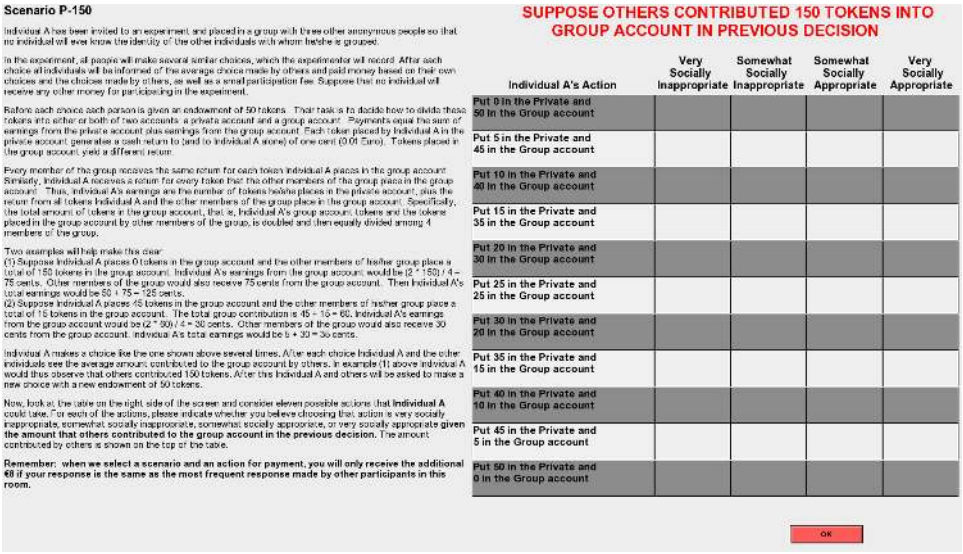


Figure D10: Repeated PG Elicitation Treatment, Pages 4-7. The four pages differ in the amount that others contributed in the previous period: 0, 50, 100, 150 tokens.

E Additional Analysis

E.1 Analysis of RF Behavior

Figure E1 displays empirical CDFs of waiting times in the RF task by treatment. Note again the strong effect of the NoRule manipulation on waiting times. Perhaps surprisingly we see also a slightly different distribution of waiting times in the Reverse treatment with more rule-breakers than in the non-reverse treatments. This suggests that there may be some spillovers from outcomes in the PG game into the willingness to follow rules in the RF task. Below, we report regression analysis of individual RF behavior.

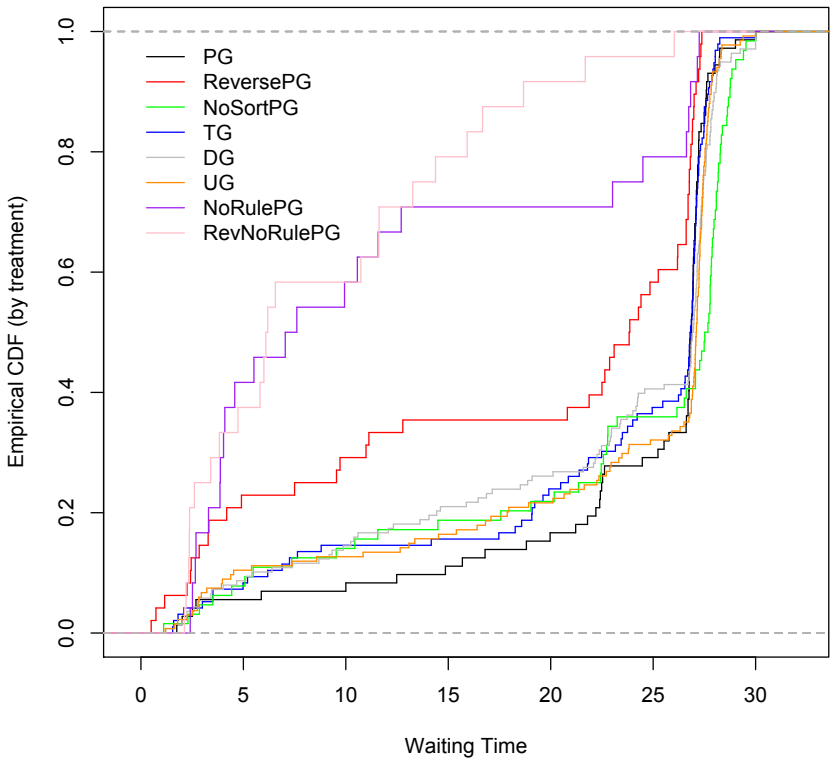


Figure E1: Empirical CDFs of RF Task Waiting Times by Treatment.

To explore sources of individual differences in rule-following proclivity, we pool the data from all experimental sessions and report regression analysis explaining RF task behavior in terms of subjects' moral foundation scores with controls for demographic characteristics and the NoRule and Reverse treatments. The dependent variable is the total time the subject spent waiting at stop lights in the RF task, and the independent variables are subjects' scores for each of the five moral values, age, gender, dummy variables for the reverse-PG and NoRule treatments, an interaction dummy between NoRule and reverse-PG, field of study dummies, a dummy for non-European subjects, and a constant term. In the reverse-PG treatment, we also control for subjects' own mean contribution to the public good as well as the mean contribution of others in their group. This allows us to test for spillovers from the PG game into the RF task. Most

of our subjects are business majors, so the field of study dummies indicate differences from the average business major. Note that we do not include a dummy for any other treatments since all other subjects were unaware of the details of the second stage when making their RF task decisions.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.650	9.170	3.015	0.003***
Reverse	0.060	3.488	0.017	0.986
NoRule	-10.293	1.412	-7.291	0.000***
Harm	-0.122	0.108	-1.131	0.259
Fairness	0.099	0.111	0.887	0.375
InGroup	-0.021	0.107	-0.192	0.848
Authority	0.014	0.104	0.131	0.895
Purity	0.089	0.093	0.951	0.342
Age	-0.003	0.004	-0.710	0.478
Female	2.550	0.790	3.229	0.001***
Non-European	0.227	1.203	0.189	0.850
Economics	-1.135	0.897	-1.264	0.207
Law	-0.862	1.646	-0.524	0.601
Psych	0.731	1.595	0.458	0.647
Other	-1.052	1.086	-0.969	0.333
Reverse_Contrib	-0.009	0.092	-0.100	0.921
Reverse_Others'_Contrib	-0.055	0.044	-1.257	0.209

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
N = 600; F-Statistic = 7.47, p-value < 0.01

Table E1: Determinants of Waiting Time, OLS Regression

Table E1 reports the estimation results. First, note again that subjects are substantially more likely to break the rule in the NoRule treatment than in the other treatments, which indicates that an explicitly stated verbal rule, with no strings attached, is sufficient to induce rule following. Second, we find that female subjects are less likely to break the rule than their male counterparts, and that age has no noticeable effect on rule breaking. Women are also less likely to cross at red lights in observational studies of pedestrian behavior in Amman, Jordan and Tel-Aviv, Israel (Hamed, 2001; Rosenbloom, 2009). In Amman, age is also negatively correlated with crossing, but because our sample consists of university students, our data may lack the variability necessary to identify an effect. Of course, age is likely to matter far less in a simulated environment because it no longer correlates with the ability to quickly cross the road. Our field-of-study dummies and the non-European dummy are insignificant. Surprisingly, none of the moral values scores are correlated with rule-following; we expected, at the least, that the instrument measuring respect for authority would correlate with RF-task behavior. Finally, after controlling for experience in the PG treatment and other covariates, we find no evidence of an effect of the Reverse treatment or of any statistically significant spillovers.

E.2 Additional Analysis of the PG treatment

E.2.1 Non-parametric Tests

In this subsection we provide additional support for the findings in Section 4.2 using non-parametric tests. In support of the findings in Figure 3, Table E2 reports period by period Wilcoxon runk-sum test comparisons of the contributions.

Period	1	2	3	4	5	6	7	8	9	10
Test Statistic $W_{9,9}$	40	47	61.5	55.5	63	64	60	60.5	61	67
`p-value`	.53	.30	**.03**	.10	**.03**	**.02**	**.05**	**.04**	**.04**	**.01**

Bolded entries statistically significant with p-value < 0.05. One-sided tests.
$W_{n,m}$ indicates the Wilcoxon test statistic with n and m observations per group type.

Table E2: Wilcoxon tests of mean group contribution by period in PG treatment.

In 7 out of 10 periods, we reject the null hypothesis of equal mean contributions in favor of the alternative that rule-following groups contribute more to the public good. Furthermore, comparing average group contributions over the first 5 periods and last 5 periods, additional Wilcoxon tests indicate that mean group contribution is significantly higher in rule-following groups than in rule-breaking groups in both early periods ($W_{9,9} = 61$, p-value $= 0.039$, one-sided test) and late ($W_{9,9} = 65$, p-value $= 0.017$, one-sided test).

Now we provide the support to our claims about additional robustness NoSort-PG and Reverse-PG treatments with non-parametric tests. A Wilcoxon test rejects the null hypothesis of equal mean contributions in period 10 between rule-followers and the pooled NoSort-PG and Reverse-PG groups in favor of the alternative that contributions are higher among rule-following groups ($W_{9,28} = 181$, p-value $= 0.027$, one-sided test), but we cannot reject the null hypothesis of equal mean contributions between rule-breakers and NoSort-PG and Reverse-PG groups ($W_{9,28} = 159$, p-value $= 0.257$, two-sided test). These results are essentially unchanged if we perform separate tests for the NoSort- and Reverse-PG treatments.

E.2.2 Support for Figure 4

This section provides support for the evidence in Figure 4. In a mixed-effects panel regression with random effects for each group and standard errors clustered at the group, we regress group-level mean contributions in each period on dummies for assignment to a No Sort group, Reverse group, rule-following group, or rule-breaking group. We also include a period trend and period×group-type interaction terms. The results are reported in Table E3. The results are qualitatively unchanged if we also include the NoRule groups, as in Figure E2.

PG Contribution	Coef.	Std.Err.	z-value	Pr($>\|z\|$)
Intercept (NoSort)	34.245	2.734	12.52	0.000^{***}
Rule-Following Group	–1.381	2.927	–0.47	0.637
Rule-Breaking Group	–0.593	3.038	–0.20	0.845
Reverse	–1.992	3.326	–0.60	0.549
Period	–1.292	0.361	–3.58	0.000^{***}
Rule-Following Group × Period	1.851	0.603	3.07	0.002^{***}
Rule-Breaking Group × Period	–0.623	0.53	–1.17	0.240
Reverse × Period	–0.232	0.463	–0.50	0.617

N = 460; 46 groups × 10 periods; standard errors clustered at the group level
$^{*} - (p < 0.10),^{**} - (p < 0.05),^{***} - (p < 0.01)$

Table E3: Mixed Effects Estimates Showing Treatment Effects in PG Contributions

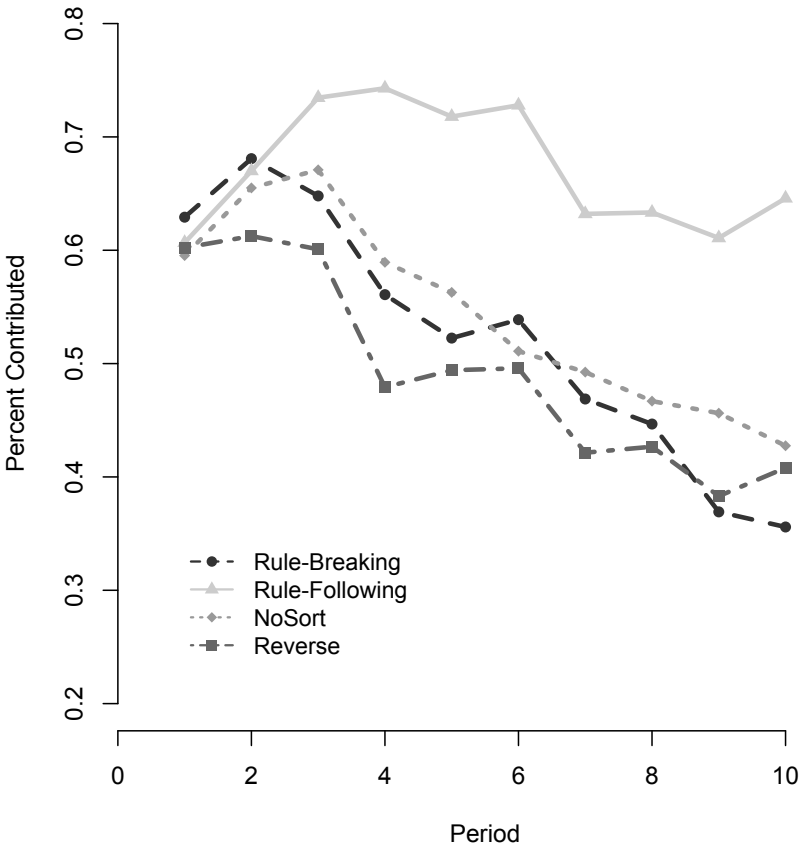


Figure E2: Time Series of Mean Group Public Good Contributions by Treatment, Rule-Following and Rule-Breaking. This figure includes sorted NoRule sessions in the computation of means for the rule-following and rule-breaking time series. While the rule-breakers mean is slightly higher than when the NoRule sessions are excluded (compare Figure 4), this is no surprise since the NoRule treatment classifies as rule-breakers many of those who would have followed the rule had we stated it expressly. Moreover, clear differences between types remain.

A positive and significant intercept and the absence of significant group-type dummies indicate no differences in initial contributions across group types. A negative and highly significant coefficient on the period trend indicates that contributions decline significantly over time in the NoSort groups, and negative, but insignificant coefficients on the period interactions with Reverse and Rule-Breaking Group dummies indicate that contributions also decline over time in those groups (Wald tests confirm that the sum of the period and Reverse×period terms differs from zero and that the sum of the period and Rule-Breaking×period differs from zero, both p-values < 0.01). On the other hand, a positive and highly significant coefficient on Rule-Following×period suggests that temporal decline is offset in rule-following groups, and a Wald test cannot reject the null hypothesis that the period and Rule-Following×period interaction sum to 0 (p-value $= 0.42$). Group-level data is inconsistent with the hypothesis that the RF task measures prosocial preferences, and it provides further support for Finding 1.

The next table reports our estimates of the impact of RF task behavior on contributions to the public good in our unsorted treatments. A positive and significant coefficient on waiting time would indicate

that our RF task inadvertently measures the strength of other-regarding preferences. However, we find no evidence for this interpretation (see Section 4.2).

PG Contribution	Coef.	Std.Err.	z-value	Pr(> |z|)
Intercept	34.794	2.726	12.76	0.000^{***}
Period	–1.392	0.244	–5.71	0.000^{***}
Time Waited (ϕ_i)	–0.043	0.121	–0.36	0.720

N = 1120; 28 groups × 4 subjects × 10 periods
Standard errors clustered at the group level
$^{*} - (p < 0.10),^{**} - (p < 0.05),^{***} - (p < 0.01)$

Table E4: Mixed Effects Estimates of the Relationship between RF Task Behavior and PG Contributions in the NoSort and Reverse Treatments

E.3 Additional Analysis of the TG Treatment

E.3.1 Regression Analysis on Amount Sent

The next table reports mixed effects estimates of the amount sent by rule-following and rule-breaking groups in the TG treatment. Each observation is one group's mean amount sent in a period. We regress amount sent on an intercept, a period trend, a rule-following group dummy and a period×rule-following interaction. We include random effects for each group and cluster standard errors at the group level. Insignificant coefficients on all but the intercept indicate no difference in behavior across group types.

TG Amount Sent	Coef.	Std.Err.	z-value	Pr(> |z|)
Intercept (Rule-Breakers)	36.57	5.697	6.42	0.000^{***}
Period	1.344	2.533	0.53	0.596
Rule-Following Group	0.028	9.306	0.00	0.998
Rule-Following Group × Period	–0.177	3.267	–0.05	0.957

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

Table E5: Mixed Effects Estimates of Trust (Amount Sent) by Group Type in the TG Treatment

The following figure displays the raw data of amount received and amount returned in the TG treatment.

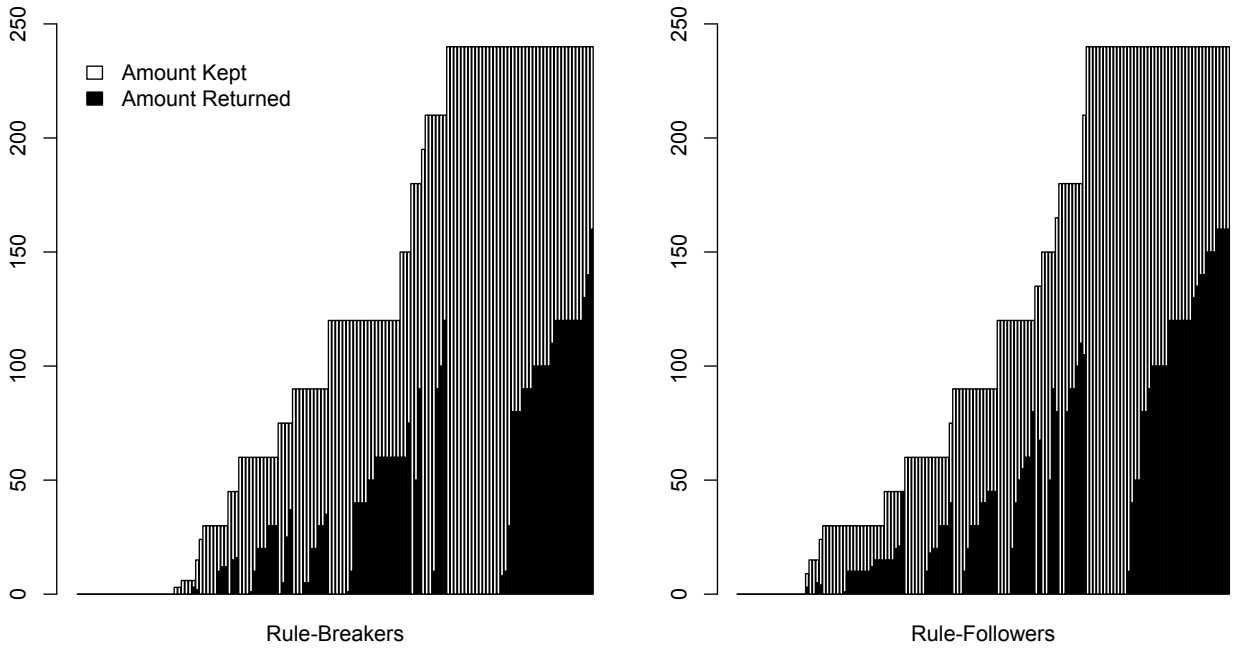


Figure E3: Amount Received, Kept and Returned in the TG Treatment, by Group Type

E.3.2 Non-parametric Tests on Returns

The regression analysis in Section 4.3 is further supported by Wilcoxon rank-sum tests of the null hypothesis of equal mean return on trust in rule-following and rule-breaking groups for each trial (1-3), where the first trial is defined as the first time a subject was in the role of first-mover, and observations are excluded where the first mover sent 0. In the first trial where all observations are independent, we weakly reject the null hypothesis in favor of the alternative hypothesis that mean return on trust is higher in rule-following groups ($W_{43,42} = 752.5$, p-value $= 0.089$, one-sided test). This is also true for the second trial ($W_{42,37} = 633.5$, p-value $= 0.071$, one-sided test); however we cannot reject the null hypothesis for the third trial ($W_{39,38} = 698.5$, p-value $= 0.331$, one-sided test).[11] Pooling the data and taking the mean return on trust for each subject over all three trials, another Wilcoxon test weakly rejects the null hypothesis of equal mean returns ($W_{48,47} = 950.5$, p-value $= 0.092$, one-sided test).

E.4 Conditional Logit Models

In this section we fit the conditional logit model of norm-dependent utility to our data in Dictator, Ultimatum, Trust and Public Goods games using the RF task estimation of ϕ and the norms elicited in separate sessions. We take the average norm elicited for each role of the player in each game assuming that the appropriateness levels are $\{-1, -\frac{1}{3}, \frac{1}{3}, 1\}$.[12] Thus our independent variables in the conditional logit regressions are the two terms of the norm-dependent utility function (see Section 2): personal payoff and the benefit/cost of following/deviating from the norm. The payoff is calculated as follows. For responders (last movers in a game), the payoff for each available action is just the payoff they receive should they

[11] The number of observations changes because we only consider cases where first movers sent a positive amount.
[12] In our procedures in this section we follow Krupka and Weber (2013).

choose that action.[13] For proposers, the payoff of an action is the *expected* payoff that proposer would get if he chose that action. The expectation is taken using the available data on responders' choices (for Ultimatum and Trust games). To calculate the possible payoff for an action in the repeated Public Goods game, we assume that the players expect others to choose the same contributions as in the last period. To compute the cost of deviation from the norm we multiply the normalized RF task score by the average norm elicited from others (for a given action).[14]

To run the conditional logit regressions for the Dictator game we expanded each observation (Dictator sending some amount of money) to 9 observations, corresponding to 9 observed choices that were made in the Dictator game by all subjects $(0, 1, 1.5, 2, 4, 5, 6, 7, 8)$. We generated a new binary variable (dependent variable in conditional logit regression) which was equal to 1 for the actual choice of the Dictator and 0 for the other 8 choices (1:8 matching). The payoff for each of the 9 actions (independent variable) was equal to the amount the Dictator left for herself. The cost of deviation from the norm was calculated for each of the 9 actions by multiplying the normalized RF task score for given subject by the average average elicited norm (Figure 8d). Since we elicited the norm only for some offers $(0, 2, 4, 8, 10, 12, 14, 16)$ we calculated the norm for other offers by linearly interpolating the closest values that were elicited (e.g. for the offer of 5, the norm was calculated as the average of elicited norm for 4 and 6). In the conditional logit regressions the data were grouped by subject with robust errors.

An analogous procedure was used for the proposers' and responders' choices in the Ultimatum game. For proposers each observation was expanded to 8 observations (1:7 matching), corresponding to all observed choices of proposers $(1, 2, 4, 5, 6, 7, 8, 9)$. The expected payoff of the proposer for sending $€x$ was calculated as $(16-x)*\text{Prob}(\text{Accepted}|x)$, where $\text{Prob}(\text{Accepted}|x)$ was estimated from the existing data on responders' rejection thresholds. The cost of deviation from the norm was calculated as the subject's RF task score times the average elicited norm (Figure 8e). As in the Dictator game, linear interpolation was used to calculate the norm for offers that were observed but not directly elicited. For the responders each observation was expanded to 9 observations (1:8 matching) corresponding to the 9 observed choices of rejection threshold $(0, 1, 2, 3, 4, 5, 6, 7, 8)$. Since the responders did not choose a particular action, the expected payoff of each of the 9 thresholds was calculated using the existing proposers' offers. The payoff for threshold x (accept any offer above or equal to x) is $\sum_{y=x}^{16} \text{Prob}(y)y$, where $\text{Prob}(y)$ is the estimated probability of receiving offer y. The cost of deviation from the norm was calculated as the subject's RF task score times the average elicited norm (Figure 8f). In the conditional logit regressions the data were grouped by subject with robust errors.

Table E5 shows the output of the conditional logit regressions for the Dictator game and for proposers and responders in the Ultimatum game. One can see that in the Dictator game the probability of choosing a particular action is proportional to the personal payoff that the action brings and the social appropriateness of this action with the individual RF task score taken into account. Thus, in the Dictator game, the sensitivity to the cost of deviation from the norm g increases with our estimate of ϕ, the RF task score.

The results for the Ultimatum game are less clear. We see that the choices of proposers are influenced by the expected payoff but not by the norm. This is consistent with our predictions that proposers do not always follow the norm, but rather act strategically given their belief about the ϕ parameter of the responder. Unfortunately, we cannot say much about the responders in the Ultimatum game. This may be the result of too much noise in the data.

[13] Except for the Ultimatum game where responders chose a rejection threshold. Here we estimated the expected payoff given the available proposers' offers.

[14] Normalization of the RF task score is just the number of seconds spent waiting at the traffic lights divided by 25, so that rule-followers have ϕ equal to 1 or higher.

	clogit: Pr(Choice)					
	DG	UG Prop	UG Resp	TG Prop	TG Resp	PG
payoff	0.269**	1.335***	–0.209*	–0.029**	0.004***	–0.002
	(0.132)	(0.477)	(0.124)	(0.013)	(0.001)	(0.003)
ϕg	1.538***	–0.206	–0.915	–0.257	–1.234***	–0.561
	(0.591)	(0.716)	(3.520)	(0.365)	(0.355)	(0.351)
highgr×ϕg				0.050	0.726*	1.460**
				(0.524)	(0.423)	(0.636)
N	603	552	621	3456	2651	7128
Num. of subj.	67	69	69	96	96	72
Independent N	67	69	69	24	24	18

Table E5: Conditional logit regressions of choices in the Dictator, Ultimatum, Trust, and Public Goods games. In all regressions data are grouped by subject. In Dictator and Ultimatum games errors are robust. In Trust and Public Goods games errors are clustered by group. We exclude observations for Trust game responders when proposers send 0.

We analyzed the data for the Trust game in a similar fashion. For the proposers, each observation was expanded to 12 observations (offers of 0, 10, 15, 20, 25, 30, 40, 50 ,60 ,70, 80). Several offers in the data had to be rounded, in order to decrease the number of available actions with only 1 observation. For each action the expected payoff of the proposer was calculated as follows. For offer x the expected payoff was $80 - x + \text{E}[\text{amount returned}|x]$, where $\text{E}[\text{amount returned}|x]$ is the expected amount returned given offer x as estimated from the responders' data. The cost of deviation from the norm for each of the 12 actions was calculated by multiplying the proposer's RF task score by the elicited norm (Figure 8b) with linear interpolation where necessary. For responders we elicited 5 norms conditional on receiving different amounts from the proposer: 3, 60, 120, 180, 240 tokens. Given that we did not find significant differences among these 5 conditional norms (Figure 8c), we used the average of them to calculate the cost of deviation from the norm. For responders we calculated the percentage of the amount received that was sent back. Each observation was expanded to 11 observations (proportion sent back: $0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$). The real percentages in the data were rounded to nearest single number after decimal point. The payoff of the responders was known and equal to 80 tokens plus the amount received (3 times the amount sent) minus the amount sent back. In the conditional logit regressions the data were grouped by subject (3:33 matching for proposers; 3:30 matching for responders) with errors clustered at the highest level of nested clusters (groups of 4).

Table E5 shows the results of fitting the conditional logit model to the data for the Trust game. Unlike in the Dictator and Ultimatum games, in the Trust game we had groups sorted according to the scores in the RF task. Thus, in the logit model we used the variable `highgr`, which is 1 if the group of subjects are assortatively matched rule-followers and 0 otherwise.[15] We see no significant effect of the norm on proposers, which is consistent with our interpretation that proposer behavior is strategic. For responders, the results indicate that there is a negative weight 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.

[15] Here we interact highgr only with the ϕg term of the utility, since under the utility specification we've chosen, rule-followers and rule-breakers do not have different utility of money. Adding interaction highgr×payoff leaves the results qualitatively unchanged.

For the Public Goods game we looked at all choices of all subjects in periods 2 to 10. We disregarded the time dimension and considered the dependency of the probability of the choice of contribution on the contributions of others in previous period. Each choice was expanded to 11 observations (contribute to group account: $0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50$). Several observations had to be rounded to one of these values in order to avoid having the action with only one observation. Since each period is a simultaneous move game, we needed to find an estimate for each subject's expected payoff in this period. We decided to compute the expected payoff for contribution x as $50 - x + 0.5 * (y + x)$, where y is the sum of contributions of other 3 members of subject's group in the previous period.[16] To calculate the cost of deviation from the norm we used the conditional norms elicited for the Public Goods game (see Figure 8a). So, for each observation and each available action the norm part of the utility was estimated as the subject's score in the RF task times the social appropriateness of the action as determined by the norm *dependent* on the sum of contributions of others in the previous period.[17] Since each subject made 9 choices in this data set for the conditional logit we grouped the data by subject (therefore having 9:90 matching) and clustered the errors by the highest nested cluster (group of 4).

Table E5 shows the results. Here as in Trust game regression variable `highgr` is 1 for the groups of rule-followers and 0 for the groups of rule-breakers. We see that the groups of rule-followers react significantly more positively to the norm than the groups of rule breakers (interaction with `highgr` is significant, $p = 0.022$). The overall effect of the norm on the probability of choosing the actions in rule-following groups is 0.899 ($p = 0.104$). Thus, the effect of the norm on the choice of rule-followers is marginally significant.

Our results demonstrate that in rule-following groups subjects increase the probability of choosing an action if the normative cost of not following it decreases. The insignificance of the coefficient on the payoff shows that following the norm of conditional cooperation is the main factor shaping decisions in repeated Public Goods game.

Taken together, the results of our conditional logit estimates provide additional support for our interpretation of the data. Rule-followers tend to be more influenced by norms than rule-breakers in Dictator and Public Goods games.

References

Andersen, Steffen, Seda Ertaç, Uri Gneezy, Moshe Hoffman and John A. List (2011), "Stakes Matter in Ultimatum Games." *American Economic Review*, 101(7), 3427–39.

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.

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.

[16] This is a rather ad hoc assumption. However, the model with expected payoff fits much better than the one where payoff is just the amount left in the private account.

[17] The conditional norms were elicited for others' sum of contributions equal to 0, 50, 100 and 150. Thus, in order to determine which actual norm should be used for each case we rounded the others' contributions to one of these levels.

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

Engel, Christoph and Michael Kurschilgen (2013), "The "Jurisdiction Of The Man Within": Intrinsic Norms in a Public Goods Experiment.", mimeo, Max Planck Institute for Research on Collective Goods.

Engelmann, Dirk and Martin Strobel (2004), "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments." *American Economic Review*, 94(4), 857–869.

Fehr, Ernst and Urs Fischbacher (2004), "Third-party punishment and social norms." *Evolution and human behavior*, 25(2), 63–87.

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

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

Hamed, Mohammed M. (2001), "Analysis of pedestrians' behavior at pedestrian crossings." *Safety Science*, 38(1), 63–82.

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.

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.

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

Osborne, Martin J. and Ariel Rubinstein (1994), A Course in Game Theory. Cambridge, Mass.: MIT Press.

Rosenbloom, Tova (2009), "Crossing at a red light: Behaviour of individuals and groups." *Transportation Research Part F: Traffic Psychology and Behaviour*, 12(5), 389–394.