Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Beliefs also make social-norm preferences social[☆]

Michael McBride[a,*], Garret Ridinger[b]

[a] Department of Economics and Experimental Social Science Laboratory, University of California, Irvine, 3151 Social Science Plaza, Irvine, CA, 92697-5100, USA
[b] Department of Managerial Sciences, College of Business, University of Nevada, Reno, 1664 N Virginia St., Reno, NV, 89557, USA

ARTICLE INFO

Article history:
Received 7 June 2019
Revised 4 June 2021
Accepted 23 September 2021
Available online 11 October 2021

Keywords:
Norms
Beliefs
Cooperation
Pro-sociality

ABSTRACT

A growing body of research reveals that various pro-social behaviors result from a desire to follow social norms. Indeed, a recent study by Kimbrough and Vostroknutov (2016) introduced the Rule-following (RF) Task and finds that an individual's willingness to follow rules in the RF Task predicts her pro-social behavior across many experimental settings. We conduct four experimental studies that use the RF Task. We find that an individual's willingness to follow rules depends on her belief about others' rule following and not just an individual-level fixed trait for norm compliance. We discuss the implications of our results for our larger understanding of human pro-sociality.

1. Introduction

Economists have traditionally emphasized the importance of repeated interaction for sustaining pro-social behavior, yet new theories explain pro-social behavior in a wider class of interactions.[1] One of these theories grounds pro-sociality in a disposition to follow social norms.[2] Variation in social norms across settings can account for how subtle changes in experimental context may dramatically change subjects' behavior (List, 2007; Dreber et al., 2013; Ridinger, 2018), and experimental evidence reveals that various behaviors do depend, at least in part, on a willingness to follow social norms (Kessler and Leider, 2012; Schram and Charness, 2015).

Attempts to incorporate a disposition to follow social norms into our understanding of decision making have emphasized two key considerations. First, an individual's disutility from violating a social norm increases in the size of the deviation from the norm, and, second, the magnitude of that disutility varies from person to person. An example of a utility function with these two elements is

$$u_i(s_i, s_{-i}) = \pi_i(s_i, s_{-i}) - k_i g(|s_i - \widehat{s}|), \tag{1}$$

[☆] We thank the Experimental Social Science Laboratory at UC Irvine for use of its facilities. We also thank colleagues and seminar and conference participants for comments and feedback. We acknowledge financial support from the University of California, Irvine.

[*] Corresponding author.
E-mail addresses: mcbride@uci.edu (M. McBride), gridinger@unr.edu (G. Ridinger).

[1] See Ridinger and McBride (2020) for a survey of recent explanations that center on preferences and beliefs. Another class of explanations focuses on the role of reputations (e.g. Benabou and Tirole, 2006; Andreoni and Bernheim, 2009).

[2] Elster (1989) cites two key features of norms. One is that they concern actions and behaviors rather than outcomes. The other is that they require some amount of collective agreement that they are norms. The economics literature on norms is large and growing, and we provide only a sampling of the most relevant works in our bibliography.

https://doi.org/10.1016/j.jebo.2021.09.030
0167-2681/© 2021 Elsevier B.V. All rights reserved.

where $\pi_i(s_i, s_{-i})$ is i's pecuniary payoff, $\widehat{s}$ is the social norm that "ought" to be followed, $k_i \geq 0$ is i's norm salience that reflects the magnitude of the disutility from violating the norm, and $g(0) = 0$ and $g' > 0$. Variants of this social-norm utility function have been used in several studies (Lopez-Perez, 2008; Kessler and Leider, 2012; Krupka and Weber, 2013; Gachter et al., 2017),[3] and more significant modifications have also been proposed to account for a wider range of behavioral patterns and experimental results (Andreoni and Bernheim, 2009; Dreber et al., 2013).

Kimbrough and Vostroknutov (2016) (KV hereafter) appeal to this utility function in their seminal study. They explain that some individuals are more willing than others to follow rules and that this variation in rule-following disposition explains differences in pro-social behavior across individuals, i.e., individuals who follow rules are more pro-social because they are more inclined to comply with the pro-social rules that are social norms. The parameter k_i in the utility function thus reflects a *general disposition to follow rules* rather than a *specific disposition to follow social norms*. KV introduce the Rule-following (RF) Task (described in detail below) in which the experimental subject is asked to follow a rule that is costly to herself without directly benefiting or harming anyone else. KV find that those who follow rules in the RF Task are also more pro-social in the Public Good, Trust, Ultimatum, and Dictator Games, which is evidence that the willingness to follow rules makes preferences social.

We argue and provide experimental evidence that it is not just a disposition to follow rules that makes preferences social; in particular, believing that others will follow the rules also makes preferences social. A key implication of this clarification is that an individual who follows a rule in one setting might not follow the rule in another setting if she believes that the rates of rule following by others differ across those settings. Thus, the correlation between following rules and acting pro-socially found by KV can be broken by manipulating beliefs about others' behavior. We use KV's innovative RF Task and draw from other literature on social norms in obtaining our findings. To keep our language clear, we will refer to the general observance of rules as *rule following* and the subset of rule following that consists of observing social norms as *norm compliance*.

That pro-sociality depends on beliefs about others' norm compliance has been noted before, e.g. Bicchieri (2006) claims that norm compliance increases in the proportion of the population believed to follow the social norm (p. 11). Let $\beta_i \in [0, 1]$ represent the proportion of the population believed by i to follow the social norm $\widehat{s}$, and consider this modification of utility function (1):

$$u_i(s_i, s_{-i}) = \pi_i(s_i, s_{-i}) - k_i(\beta_i) g(|s_i - \widehat{s}|, \beta_i), \tag{2}$$

$$g(|s_i - \widehat{s}|, \beta_i) = \begin{cases} g(|s_i - \widehat{s}|), & \text{if } \beta_i > \widehat{\beta}, \\ 0, & \text{if } \beta_i \leq \widehat{\beta}, \end{cases} \tag{3}$$

where $k_i(\beta_i) \geq 0$ reflects how i's norm salience depends on others' norm compliance with $k_i' > 0$, and $\widehat{\beta} \geq 0$ is the minimum rate of conformity needed for i to consider $\widehat{s}$ to be the social norm. For each i with this utility function there will exist a $\beta_i^* \geq 0$ such that i will follow the norm if and only if $\beta_i \geq \max\left\{\beta_i^*, \widehat{\beta}\right\}$.[4] Observe how this utility function captures the important distinction between the socially-approved action $\widehat{s}$ that *ought* to be taken, called an *injunctive norm*, and the belief β_i about the behavior that is actually prevalent in the community, which reflects the *descriptive norm* (Cialdini et al., 1990).[5] It also demonstrates two channels by which beliefs about others' norm compliance may affect one's preferences. First, whether an action $\widehat{s}$ is considered a norm requires that there be a sufficiently high prevalence of the behavior in the population. Second, given that $\widehat{s}$ is considered the norm, individual i's norm sensitivity $k_i(\beta_i)$ increases in the proportion that others are believed to be following the norm so that the utility loss from violating a norm increases as β_i increases.[6]

Several experimental studies confirm that social-norm compliance is conditional on the belief about others' compliance (Cialdini et al., 1990; Bicchieri, 2008; Krupka and Weber, 2009; Bicchieri and Xiao, 2009).[7] Importantly, these studies examined norm compliance but not the more general rule following, nor did they relate conditional rule following to conditional norm compliance. Indeed, recognizing the role of beliefs about others' compliance adds additional insight into KV's results. First, the RF Task does not capture a fixed rule-following salience alone but instead measures a combination of individual's

[3] A similar functional form has also been used by Heath (2008) in philosophy to represent rule-following behavior.

[4] We derive β_i^* for a simplified linear utility case in the Appendix.

[5] Several other distinctions have been also proposed. For example, Bicchieri (2006) further partitions injunctive norms into social and moral norms, the latter being rules that are not to depend on others' behavior, and Elster (2009) offers social norms, moral norms, and quasi-moral norms as categories and directly implicates a role for emotions. Whereas in much of the literature injunctive norms are typically thought to correspond to social dilemma games, conventional norms (conventions for short) correspond to rules and patterns of behavior in coordination games. There also exist personal rules regarding behaviors that do not affect others. Finally, we note that if we restrict attention to injunctive norms, then β_i can be interpreted as i's belief about the share of the population that consider $\widehat{s}$ to be the appropriate action as defined in Krupka and Weber (2013).

[6] Discussion of both channels is implied in the literature, though we have not seen a utility function that explicitly accounts for both. For example, when identifying the conditions for a social norm to exist in a population (in her Chapter 1), Bicchieri (2006) distinguishes between the existence of a social norm (our second channel), which requires a sufficient subset of the population to be following the norm, and an individual's own conditional following of the norm (our first channel) that also requires a sufficient subset to be adhering to the norm. Interestingly, the utility function that Bicchieri provides (see Chapter 1 Appendix) does not explicitly include beliefs. We believe the utility function in Eqs. (2) and (3) is the first to explicitly include both channels.

[7] Conditional compliance is found in a wide range of experimental settings. An example is Kamei (2014), who finds in a voluntary-contribution setting that punishments increase in the number of others that punish. Bicchieri and Xiao (2009) provide evidence of the second of the two channels in a dictator game in which information about others' giving is provided.

beliefs about others' rule following and her rule-following tendency. Second, the disposition to follow rules across settings is due, at least in part, to correlation in beliefs about others' rule following across settings and not just a fixed willingness to following rules.

We present four experimental studies that confirm these suspicions. We find causal evidence that one's belief about others' rule following matters fundamentally for rule following, just as it does for norm compliance. In each study, the subject completes the RF Task and the Dictator Game with proceeds going to a charity of the subject's choice. Our work does not determine which of the two channels mentioned above is operating, and either or both could be in effect. Our focus is on providing causal evidence that beliefs affect compliance.

Study 1 adds belief elicitation to the RF Task and the Dictator Game. We replicate KV's finding that rule following in the RF Task is correlated with higher giving in the Dictator Game. We also report a new finding that both rule following in the RF Task and giving in the Dictator Game are positively correlated and predicted by beliefs about other subjects' rule-following and giving. Although the result is correlational not causal, it lends credibility to our conjecture that beliefs may be causing both rule following in the RF Task and norm compliance in the Dictator Game. Studies 2–4 obtain causal evidence.

Study 2 introduces a strategy-method version of the RF Task which requires subjects to choose different rates of rule following for different levels of rule following by other subjects. We find that most subjects choose to follow the rule at higher rates when others follow the rule at higher rates, revealing that rule-following behavior is not fixed across settings but is instead highly conditional on others' behavior.

Study 3 replicates Study 1 but with the addition of exogenous shocks to subjects' beliefs about the level of others' rule following in the RF Task. Before completing the RF Task, half of the subjects receive a signal that the proportion of others that follow the rule in the RF Task in a prior experiment was high, and the other half of the subjects receive a signal that others' rule following in the prior experiment was low. Those that receive the high signal follow the rule at a significantly higher rate than those that receive the low signal, thus providing causal evidence that beliefs about others' rule following affects their own rule following. Moreover, the correlation between rule following in the RF Task and donating in the Dictator Game within subjects disappears as a result of the exogenous shock to beliefs in the RF Task. Subjects who received the negative shock to their beliefs waited at fewer lights but donated just as much as the subjects that waited at many lights. This last finding reveals that the link between following rules and complying with norms found by KV depends on the beliefs in others' following and compliance and not just a fixed salience for following rules.

Study 4 replicates Study 1 but with the addition of exogenous shock to subjects beliefs about others' donations in the Dictator Game. Before completing the Dictator Game, half of the subjects receive a signal that earlier subjects making the same decision made large donations, and the other half of the subjects receive a signal that others' donations were small. Those that received the large-donation signal donate at significantly higher levels than the others, indicating that an increase in the believed conformity of others increases one's own conformity to the social norm.

Our paper's foremost contribution is causal evidence that following rules, like complying with norms, is belief conditional. We show that beliefs about others' rule following—not just a rule-following propensity alone—also makes preferences social. We replicate KV's finding that pro-sociality is associated with a disposition to follow rules, but we further show that the link depends on beliefs about others' rule following. A further implication is that rule following across settings depends on beliefs and information available about others' rule following in those different settings.

Our paper thus contributes to multiple literatures, including that which seeks to understand human pro-sociality (Bowles and Gintis, 2011), that which examines the role of social learning and imitating in the spread of behaviors (Boyd and Richerson, 1988; Richerson and Boyd, 2008), and that, mentioned earlier, which seeks how to best model the role of social norms in individual decision making. Another related literature considers peer effects that create anti-social and pro-social behaviors. Gino et al. (2009) find that unethical behavior increased when unethical confederates were from the in-group member but decreased when from the out-group. When individuals have the opportunity to lie to increase their own personal payoff, studies have found that knowledge of others' dishonest behavior increases dishonesty (Kroher and Wolbring, 2015; Diekmann et al., 2015; Brunner and Ostermaier, 2019). Thoni and Gachter (2015) conduct a gift-exchange game in which individuals can revise their efforts after learning about the effort of another player, and they find that subjects lower their efforts toward their peers but do not raise them. Dimant (2017) finds that closer social-distance and knowledge of others' behavior has a greater impact on increasing anti-social behavior than on pro-social behavior. Charness et al. (2019) find "opportunistically conformist" behavior such that subjects were more likely to conform if doing so increased their own expected payoffs. In general, approaches in which preferences are outcome based, like that innovated by Fehr and Schmidt (1999), have been highly influential in understanding experimental research. Our experimental manipulations in Study 3 and 4 contribute to these literatures by examining whether one's own behavior is affected by information about others' behavior.

There is a healthy scholarly debate about which framework—outcome-based utility with peer effects or inequity aversion, a norm-based utility, or something else entirely—is the best framework for understanding social behavior (Ridinger and McBride, 2020). For example, Gachter et al. (2013) conclude that both social-norms and inequity aversion can explain peer effects in the gift-exchange game. Notably, they find that inequity aversion better predicts behavior within a given treatment but does worse explaining differences across treatments. Lahno and Serra-Garcia (2015) also provide evidence that peer effects on risk preferences are best explained by both distributional concerns like inequity aversion and non-distributional concerns such as social norms. Still other experimental work has found that a social-norm approach best explains behavior in Dictator Games (Fershtman et al., 2012; Krupka and Weber, 2013). We acknowledge that the pro-social behavior in our

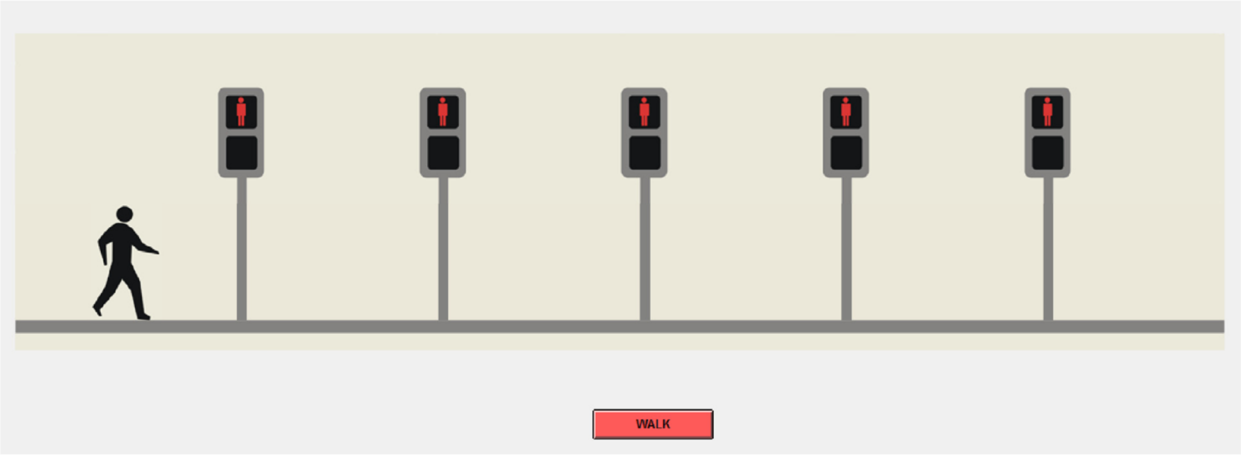

Fig. 1. Rule-following Task Screen.

experiment may potentially be explained by multiple frameworks, and we are not seeking to resolve this debate. However, we are motivated by the literature on social norms such as KV's innovative work, so we use the norm framework.

We know of only one other economics experiment that examined the conditional nature of rule following. Desmet and Engel (2017) ask experimental subjects to move sliders on a computer screen. Each subject's payoff is increasing in the number of sliders moved, but there is a rule to not move more than a pre-announced number of sliders. They obtain both unconditional and (strategy-method) conditional behavior via different treatments, elicit beliefs about unconditional choices, and find that rule compliance is increasing in the belief about others' compliance. Our project differs in significant ways. First, we use the RF Task that has previously been used by KV to measure preferences for rule following. Second, because of the potentially confounding presence of an experimenter-demand effect in the strategy method, we conduct additional treatments (Studies 3 and 4) that exogenously shock beliefs. Third, we directly connect the preference for rule-following behavior to pro-social behavior. Finally, we show that the apparent correlation between rule following and pro-sociality across settings can be broken by manipulating beliefs about others' behavior.

2. The rule-following (RF) task

In KV's RF Task, the subject controls how quickly a human stick figure walks from left to right across the computer screen. See Fig. 1 (and the supplemental appendix). To cross the screen, the figure must pass five light signals. Each light signal begins red, and when the figure arrives at a signal, the figure automatically stops. The light remains red for five seconds before turning green. The figure remains stopped at the signal until the subject clicks a button labeled WALK. After clicking WALK, the figure will walk past the signal no matter the color of the signal and will continue walking until the next signal. To complete the task, the subject must therefore click WALK once at each signal (clicking WALK while the figures is already walking from signal to signal has no effect). Importantly, the subject can click WALK even while the signal is still red, and the subject crosses the screen most quickly by walking immediately rather than waiting for lights to turn green.

The subject begins with a $7 endowment. This amount decreases by $0.07 every second, so the longer the subject takes to cross the screen, the less money the subject receives as payment.[8] The payment is based solely on the amount of time taken to cross the stick figure across the screen. The payment does not depend on whether the subject waited at red signals.

The instructions explicitly communicate a rule with the following text displayed on the screen: "The rule is to wait at each stop light until it turns green." As KV explain (pp. 615–616):

> 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 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.

[8] KV used an *e* 8 endowment, with a decrease of *e* 0.08 per second.

Importantly, the subject's decision in the RF Task affects her own monetary payment but not the monetary payment for any other subject. The RF Task intends to capture the subject's desire to follow the rule independently of how violating the rule might harm another subject. This feature distinguishes the RF Task from a task in which the subject's action financially benefits another subject at a cost to herself.

KV also ran diagnostic treatments in which the explicit statement of the rule was removed. The subject must then determine how to act without any explicit guidance about a rule. Waiting times decreased dramatically in those diagnostic sessions. Their study does not determine exactly why, and there are competing explanations because stating the rule may serve two different purposes. One is that stating the rule creates knowledge of the rule, but another is that it primes subjects' beliefs about how many others will follow the rule. Notice the distinction between, first, unawareness of a norm and, second awareness of norm but an imperfect belief about the rate of others' conformity. The lower rate of rule following in the diagnostic session may be due to one or other. Our experimental design bears this distinction in mind. By stating the rule explicitly in the RF Task, we ensure that subjects know about the rule, and we can therefore manipulate beliefs about others' rates of rule following.

3. Study 1: belief elicitation

3.1. Study 1 procedures

Study 1 ($n = 62$) consists of three parts.[9] The first part is the Dictator Game with a charitable organization. The subject is given $5 and told that she will get to choose how much of the $5 to donate to a charitable organization of her own choice. The amount selected will be donated by the experimenter on the subject's behalf, and the subject's name will not be used when making the donation. The subject reads a paragraph for each of four charities: Amnesty International, United Nations Children's Fund (Unicef), Doctors Without Borders, and the American Cancer Society. Each paragraph is a written description of the charitable organization's purpose, and below each paragraph is photograph of a beneficiary of the organization. The subject then selects one of the four organizations to receive her donation. Next, the subject selects an amount in whole dollars to donate to the organization she selected. The subject can choose to donate between $0 and $5.

After making her donation, the subject is asked to report what she believes was the average donation by the other subjects that selected Amnesty International, the average donation by others to Unicef, and so on. Decimals are allowed. If the subject's answer is within 0.25 of the actual average, then she receives $1 on top of what was kept (i.e., not donated) in the Dictator Game. If the answer is not within 0.25 of the true average, then the subject receives no additional money. This belief elicitation does not incentivize truthful reporting when a subject's actual belief is within 0.25 of the $0 and $5 boundaries, but it is easy for the subjects to understand and has been shown to be an effective way to elicit beliefs. Several subjects did in fact report beliefs within 0.25 of the boundaries. We cannot claim with full confidence that their reported beliefs near the boundary represent their true beliefs with perfect accuracy, however this procedure does incentivize a reported belief that is very close (within 0.25) to their true belief when the true belief is near the boundary.[10] We elicit the four beliefs separately because we expect that it is the belief that corresponds to donations made by others to the same charity that is the relevant belief reference.[11]

The second part of the study is the RF Task with belief elicitation. After completing the RF Task as described above, the subject is asked, "What do you think is the average number of lights that the other people in the room waited at?" The subject answers by entering a number via the keyboard. As with the Dictator Game belief elicitation, the subject is paid $1 for an answer within 0.25 of the true average and $0 otherwise. We note that spillover from the belief elicitation to the decision task should be minimal. First, we elicit the beliefs for the Dictator Game and RF Task on separate screens after the decisions are made, so the subjects are unaware of the belief elicitation when making their decisions. Second, even if there was concern that they could be aware, we used a belief payment that has much smaller stakes than the decision payments, which, as Blanco et al. (2010) explain, can help mitigate spillover.[12]

Our decision to elicit strategies by lights waited rather than time to finish was deliberate. Because the rule is stated in terms of waiting at lights (not in terms of time to cross the screen), the belief elicitation directly measured rule following. We also suspect that the number of lights waited at (hereafter, "the number of lights waited") is an easier number to conceptualize. There are only five signals which is a relatively low number to consider, but time to finish is a continuous variable that may be more difficult to estimate. An additional reason is that we will refer to the number of lights waited

[9] KV had the RF task done first and the experimental game done second. An exception is a treatment where the RF task was done second for the Public Good Game to test robustness. Though they did not report any meaningful difference between the two orders, we decided to do the RF task second for two reasons. First, it is a robustness check against KV's Dictator Game sessions in which the RF task was done first. Second, as a precaution against the priming of a social norm in the RF task causing a priming of a social norm in the Dictator Game. Note that the Dictator Game involves no explicit priming of a social norm. That rule-following behavior and donations in Study 1 are so similar to those in KV's data suggest that having the subjects complete the RF task first or second makes no meaningful difference.

[10] For discussion, see Charness and Dufwenberg (2006).

[11] In all subsequent analysis, we use the subject's stated belief for the chosen charity. However, we note that results are similar if we use the average reported belief about others' donations for all four charity selections.

[12] As will be seen later in the paper, the distribution of choices in our RF Task with belief elicitation in Study 1 is similar to the distribution of choices in KV's RF Task without belief elicitation. This suggests that the inclusion of the belief elicitation does not have an impact on the RF behavior.

Table 1
Summary Statistics.

	Study 1	Study 2	Study 3		Study 4		Overall
			Low	High	Low	High	
	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)
Female (Fraction)	0.65	0.76	0.71	0.52	0.63	0.74	0.71
	(0.48)	(0.43)	(0.46)	(0.50)	(0.49)	(0.44)	(0.46)
Age	20.40	20.66	20.26	20.33	20.16	19.54	20.28
	(1.62)	(1.28)	(1.77)	(1.53)	(1.57)	(1.82)	(2.09)
Number of Economic Courses	1.73	0.76	1.19	1.43	1.19	1.10	1.18
	(3.35)	(1.21)	(1.04)	(1.78)	(2.13)	(2.14)	(2.35)
Number of Statistics Courses	1.35	0.93	1.15	1.43	1.03	0.92	1.10
	(1.39)	(0.94)	(1.25)	(1.78)	(0.86)	(1.31)	(1.16)
Take Home Pay	12.53	13.5	12.75	12.55	14.05	12.72	12.93
	(1.90)	(2.00)	(1.86)	(2.23)	(1.78)	(2.00)	(2.04)
Observations	62	67	68	71	32	39	339

in Studies 2 and 3 when exogenously treating beliefs. Our terminology thus draws attention to the rule in the RF task and uses the same language for consistency across the studies.

Our decision to use only the Dictator Game was also deliberate. The Trust, Ultimatum, and Public Good Games are strategic in that each subject's own monetary payoff depends on both her own action and the action of another subject in the experiment. Such is not the case in the Dictator Game played with a charity because the subject is paired with an entity (the charitable organization) that is not an experimental subject. The Dictator Game may thus be considered the least strategic of the four games used by KV, and, consequently, it is the simplest setting to examine other-regarding behavior. Much of the recent work on social norms have focused on the Dictator Game for this reason (Bicchieri and Xiao, 2009; Krupka and Weber, 2009; Dreber et al., 2013; Krupka and Weber, 2013; Gachter et al., 2017). Our focus follows this tradition, although we acknowledge that this game does indeed have a strategic element if, as we suspect, subjects have social-norm preferences. Although monetary payoffs do not depend on others' actions in the Dictator Game, the utilities do when the utility function depends on the belief about others' conformity. Another strategic element could exist if subjects view their donations to charity as substitutes. Nonetheless, the Dictator Game remains less strategic than the other games.

Studies 1–4 were conducted in the experimental economics laboratory at a large public university in the United States. Students register to be the subject pool via an online registration system after receiving email advertisements. Days before an experiment session is scheduled, the students in the subject pool receive an email that announces our experiment session and invites them to sign up online to participate in that session. An email is sent out the night before to remind sign-ups of their scheduled session. Subjects could participate in at most one session, and there were no other exclusion restrictions for participation other than being 18 years of age.

All studies were approved by the university's Institutional Review Board (HS #2011-8378). The z-Tree software platform was used to instruct subjects and collect the subjects' choices via mouse and keyboard (Fischbacher, 2007). Upon arrival at the lab, each subject was randomly placed at one of the lab's computers. At the experiment's official starting time, the door to the laboratory was closed, and the software was started. Each study lasted about one hour. All subjects received a show-up payment of $7, plus additional earnings based on decisions made during the study. The payments were made in cash at the end of the session, after which the subject left the laboratory. Table 1 provides summary information about Study 1 and the other studies.[13] Instructions for all studies are provided in a supplemental appendix that is available from the authors.

3.2. Study 1 predictions

Because the belief-elicitation questions are asked after the subjects make their rule-following and donation decisions, the asking of the questions should not affect rule-following behavior or donation amounts. Moreover, given that the experimental procedure replicates KV's Dictator Game treatment aside from the belief elicitation questions, we expect that patterns of rule-following behavior and donations should be similar to those reported by KV.

Finally, our primary conjectures of interest are that the rates of conformity to rules and norms are increasing in the belief about the conformity of others. Observe that the utility function in Eqs. (2) and (3) can apply to either the RF Task or the Dictator Game. In the former case β_i represents i's belief about others' waiting at lights, while in the latter case β_i

[13] In addition to basic demographic information, subjects also completed the Reading the Mind in the Eyes test (RMET) (Baron-Cohen et al., 2001). These data were collected as part of a separate research project that is outside the scope of this paper. The RMET was completed by subjects in the same order in all studies, and there is no reason to believe that its inclusion in the experiment had any impact on the data of interest.

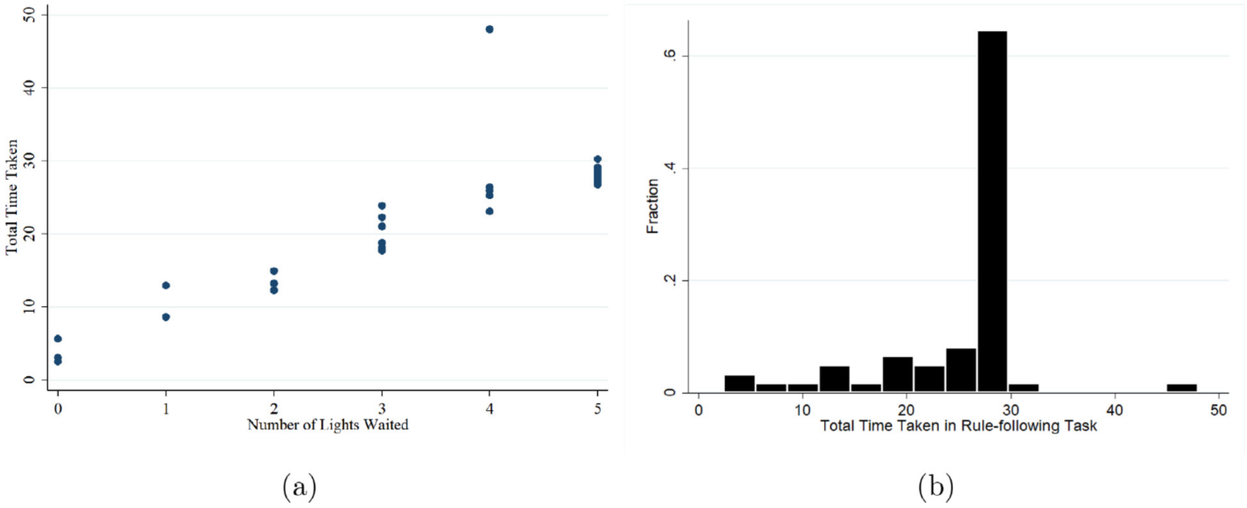

Fig. 2. Study 1: (a) Scatter plot of total time taken and number of lights waited in the Rule-following task. (b) Histogram of total time taken in the Rule-following Task.

represents *i*'s belief about others donations. In the Appendix we formally examine a linear version of the utility function in Eqs. (2) and (3) to show that an individual *i* will only follow the rule in the RF Task (or conform to the norm in the Dictator Game) if the individual's belief about others' rate of following (or conformity) is sufficiently high. Thus, following the rule will be positively correlated with the belief about others' rate of following, and conforming to the norm will be positively correlated with the belief about others' norm conformity.

Hypothesis 1:

(a) Following the rule in the RF Task will be positively correlated with the amount donated in the Dictator Game.

(b) Following the rule in the RF Task will be positively correlated with the believed average amount of rule following by others.

(c) Donation amounts in the Dictator Game will be positively correlated with the believed average amount of donating by others to the same charity.

3.3. Study 1 results

Figure 2 (a) shows that total time taken to cross the screen and the number of lights waited are strongly correlated. Thus, using either lights waited or time to finish will yield similar conclusions in our data analysis. Figure 2(b) displays the distribution of time taken. Our distribution is nearly identical to KV's, indicating that we successfully replicated their level of rule-following behavior in the RF Task.

Table 2 reports coefficients from a series of regressions. The correlation between rule following and donations is positive as predicted by Hypothesis 1(a), but it not significant. KV also found a positive but low correlation, and closer examination of their data showed that the difference in giving is largest when comparing those who follow the rule with those who never follow the rule. This pattern is illustrated in Fig. 3. As shown in Table 2 column (6), those subjects who never followed the rule always donated very little while those who waited at some or all lights donated much larger amounts, indicating that there is a strong difference in donation behavior that correlates with rule following, i.e., that the extensive margin matters more than the intensive margin. We cannot compare quartiles as did KV because the median lights waited is five. However, we do see the striking difference in donation amounts between those that waited at some or all of the lights and those that waited at zero lights.

Table 3 confirms our claims in Hypothesis 1(b)–1(c) that beliefs predict rule following and donations. An increase of one in reported belief about the average number of lights waited by others is associated with an increase of waiting at about one half more light. This effect is statistically significant at high confidence levels and economically meaningful. An increase of one light in the belief about others' rule following leads to a decrease in subject earnings by $0.17, which corresponds to a 9.7% decrease in earnings in RF the task. For donations, an increase of $1 in reported belief about the average amount donated is associated with an increase in $0.78 donated to charity. This effect is statistically significant at the 1% level.

These findings are consistent with our primary argument that both rule following and donations depend causally on beliefs about others' behavior. However, these findings are merely correlational, and we seek causal evidence in the remaining studies.

Table 2
Study 1- Predicting Donations by Rule Following.

	(1) Donation	(2) Donation	(3) Donation	(4) Donation	(5) Donation	(6) Donation
Number of Lights Waited	0.16 (0.18)					
Waited at Five Lights		0.03 (0.52)				
Waited at Four to Five Lights			-0.20 (0.59)			
Waited at Three to Five Lights				0.99 (0.74)		
Waited at Two to Five Lights					1.34 (0.91)	
Waited at One to Five Lights						2.89*** (0.32)
Female	1.05** (0.52)	1.14** (0.52)	1.18** (0.52)	1.00* (0.52)	1.07** (0.51)	0.94* (0.50)
Age	0.04 (0.13)	0.04 (0.13)	0.04 (0.13)	0.03 (0.13)	0.07 (0.13)	0.05 (0.12)
CRT	0.17 (0.19)	0.14 (0.19)	0.13 (0.20)	0.15 (0.18)	0.19 (0.19)	0.13 (0.18)
Intercept	0.97 (2.62)	1.62 (2.63)	1.76 (2.64)	0.95 (2.59)	-0.21 (2.72)	-1.29 (2.53)
N	62	62	62	62	62	62
R^2	0.101	0.085	0.087	0.118	0.125	0.206

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

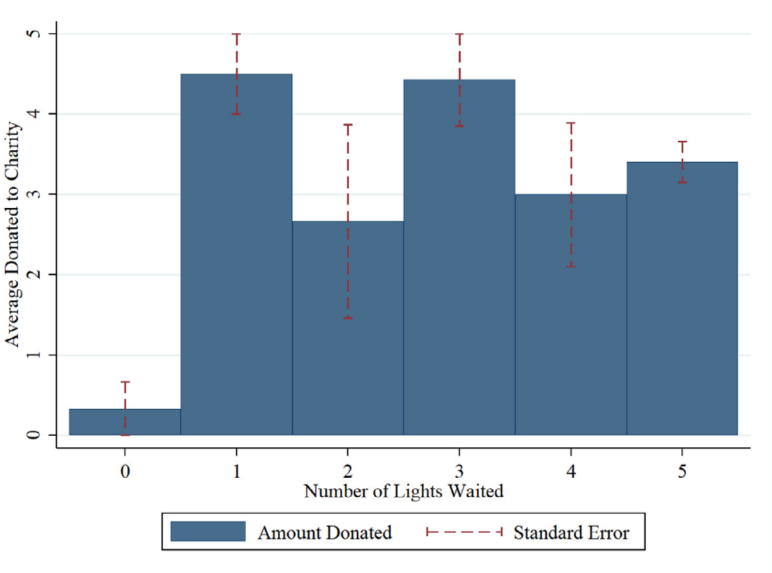

Fig. 3. Study 1- Donations To Charity by Number of Lights Waited in the Rule-following Task.

4. Study 2 strategy-method RF task

4.1. Study 2 procedures

Study 2 ($n = 67$) is identical to Study 1 except for a change in the RF Task. The RF Task is explained to the subjects, and the subjects are told that participants in a prior study completed it. Next, instead of the subject clicking the WALK button manually at each light, the subject inputs a strategy for walking that can condition on how the subjects in the prior study behaved. The computer then carries out the subject's strategy on the subject's behalf.

Specifically, the subject is told, "If the average number of lights waited by participants in the prior study was between 0.00 and 0.99, then you will be asked how many lights you want to wait at." A similar response is requested for four other scenarios, i.e., when the average was between 1.00–1.99, 2.00–2.99, 3.00–3.99, or 4.00–5.00. The subject next enters a number of lights waited for each of those five categories. After the subject enters her strategy but before the computer uses

Table 3
Study 1- Predicting Donations by Rule Following.

	(1) Number of Lights Waited	(2) Number of Lights Waited	(3) Donation	(4) Donation
Belief Other's Number of Lights Waited	0.47**	0.45**		
	(0.19)	(0.20)		
Belief Other's Donation			0.76***	0.78***
			(0.10)	(0.10)
Female		0.62		0.15
		(0.39)		(0.41)
Age		0.02		0.17*
		(0.09)		(0.10)
CRT		-0.10		0.06
		(0.15)		(0.14)
Intercept	2.47***	1.81	0.93**	-2.77
	(0.78)	(2.07)	(0.37)	(2.13)
N	62	62	62	62
R^2	0.145	0.209	0.414	0.439

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the true average to carry out the strategy on her behalf, the subject is also asked what she thinks is the average number of lights waited chosen for the first question by subjects in her same session, and so on for all five possible responses. To elicit truthful reports, we follow the same payment scheme of paying $1 for each answer that is within 0.25 of the true average. The subject is then told the actual average lights waited by the prior subjects, and her plan is carried out on her behalf by the computer using the actual average lights waited from the RF Task in Study 1.

This "strategy method" is a standard method used in experimental economics to obtain information about conditional behavior. One criticism of this method in our study is that subjects may be susceptible to a strong experimenter-demand effect that undermines the credibility of the method. Specifically, by asking the subject to report a number of lights waited for each range, we may be unintentionally signaling to the subject that she ought to enter a strategy that conditions on others' behavior. Moreover, a natural conditional, conformity strategy might be to wait at the same number of lights as the others are waiting at. This strategy-method design cannot separately identify behavior that is truly conditional on beliefs (as we suppose it is in this paper) from behavior that is conditional because of the presence of an experimenter-demand effect whereby the subject enters a conditional strategy because she believes the experimenter expects her to.

With this concern in mind, our Studies 3 and 4 offer more convincing causal evidence of the role of beliefs in following rules and social norms. Nonetheless, we believe that Study 2 is still valuable for several reasons. First, under the assumption that the experimenter-demand effect is not present, the data from the strategy method provide causal evidence about the impact of changing the belief about others' compliance on one's own compliance. Prior studies have found that direct-response methods and the strategy method yielded similar results, suggesting that experimenter-demand effects might not be present.[14] Second, the strategy-method RF data from Study 1 can be compared with data from Study1's standard RF Task to look for a similar relationship between beliefs and conformity. Finally, the strategy method is often valued because it provides detailed information about subject heterogeneity. Later in the paper we report significant differences in the types of strategies submitted by the subjects in Study 2.

4.2. Study 2 predictions

Consistent with the utility functions in Eqs. (2)-(3) and as suggested by the results from Study 1, the frequency of rule following should increase as the number of lights waited by others increases. Moreover, we conjecture that the increase in rule-following should be similar in Studies 1 and 2.

Hypothesis 2:

(a) The number of lights waited will be increasing in the number of lights waited by the subjects in the prior study.

(b) The lights waited should increase by about 0.47 for each increase of 1 in others' lights waited, thereby matching the increase in lights waited in Study 1.

4.3. Study 2 results

Figure 4 reveals that, as predicted by Hypothesis 2(a), subjects do indeed choose to wait at more lights when the average lights waited by others increases. Formal statistical analysis confirms this pattern. Table 4 displays the coefficients from a

[14] Brandts and Charness (2011) survey 29 papers that included the direct response method and the strategy method. They conclude that the majority of studies find no difference between the methods and that no study in their survey had a treatment effect in the strategy method and not in the direct response.

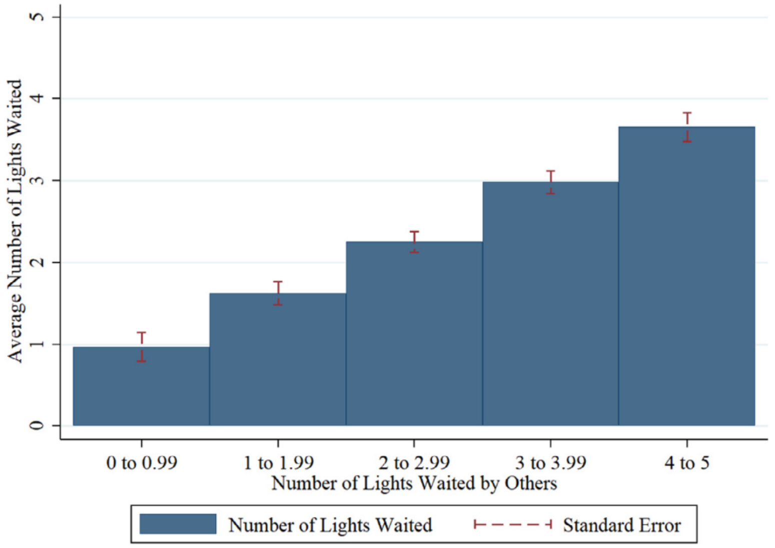

Fig. 4. Study 2- Lights Waited Given the Lights Waited by Others.

Table 4
Study 2- Predicting Number of Lights Waited by Number of Lights Waited by Others.

	(1)	(2)	(3)	(4)
	Number of	Number of	Number of	Number of
	Lights Waited	Lights Waited	Lights Waited	Lights Waited
Other's Number of Lights Waited	0.67***	0.67***		
	(0.06)	(0.06)		
Other's Number of Lights Waited (1 to 1.99)			0.66***	0.66***
			(0.08)	(0.08)
Other's Number of Lights Waited (2 to 2.99)			1.28***	1.28***
			(0.13)	(0.13)
Other's Number of Lights Waited (3 to 3.99)			2.01***	2.01***
			(0.18)	(0.19)
Other's Number of Lights Waited (4 to 4.99)			2.69***	2.69***
			(0.25)	(0.25)
Female		-0.09		-0.09
		(0.27)		(0.27)
Age		-0.03		-0.03
		(0.03)		(0.03)
CRT		0.01		0.01
		(0.13)		(0.13)
Intercept	0.95***	1.63**	0.97***	1.65**
	(0.18)	(0.65)	(0.18)	(0.65)
N	335	335	335	335
R^2	0.367	0.371	0.368	0.371

Clustered standard errors by subjects in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

series of panel regressions, with five observations per subject, one observation for each of the five belief categories. Regressions (1) and (2) use the median of each report category as the Other's Number of Lights Waited variable, an admittedly rough method of combining the beliefs into one variable that yields a single slope that measures the average increase in number of lights waited by subjects as others' number of lights waited increases by one. For example, the Other's Number of Lights Waited variable is equal to 0.5 if the Other's Number of Lights Waited category was from 0 to 0.99 and equal to 1.5 if the Other's Number of Lights Waited category was from 1 to 1.99. This increase of 0.67 shown in Table 4 is larger than the 0.47 found in Study 1, contrary to our prediction in Hypothesis 2(b). Regressions (3) and (4) reveal a similar pattern but broken up by the response categories.[15] The coefficients progressively increase category by category, indicating that an individual's number of lights waited increases in others' lights waited as predicted.

[15] We find similar results using non-parametric tests. All pair-wise two-sided Fisher-Pitman permutation tests comparing the number of lights chosen by subjects between each category of the number of lights by others are significantly different at the 0.1% level.

Although the rate of increase is of similar order of magnitude in Study 2 as in Study 1, the difference is worth consideration. We cannot determine the exact cause of the difference, but it is possible that some subjects show a larger willingness to react to the increase in others' rule following because of the aforementioned experimenter-demand effect. Another possibility is that the lower risk associated with rule-following in the strategy method leads to a higher rate of rule-following behavior in Study 2 than Study 1. Yet another possibility is that the difference is due to the differences in how the two are calculated. The Study 1 slope is calculated across subjects using that subject's mean belief, whereas for Study 2 we have several choices per subject. It may also be that there are systematic differences between those that have high beliefs and those that have low beliefs in Study 1 that lead to a different aggregate slope using only each's single reported belief.

5. Study 3: high-low RF task

5.1. Study 3 procedures

Study 3 replicates Study 1 with one change. We selected ten subjects from the prior study, put them into two groups – a "high" group and a "low" group. The subject in Study 3 is randomly assigned one of these groups of five for her reference, the first group being the HIGH treatment ($n = 71$), the second group being the LOW treatment ($n = 68$). After being explained the RF task but before undertaking the task, the subject reads on the screen: "Student participants in a prior experiment completed the same exact task. Five of these prior participants have been randomly selected, and the number of lights each waited is reported below." Subjects were not told that the groups of five were pre-arranged into a HIGH and LOW groups. Not revealing our procedure for creating these two groupings is similar to other laboratory experiments in which subjects are randomly placed into treatments without being told why or how the researcher chose the treatment conditions that are randomly assigned.[16]

For the HIGH treatment, the screen then reports:

"Person 1: waited at 5 lights."
"Person 2: waited at 3 lights."
"Person 3: waited at 5 lights."
"Person 4: waited at 5 lights."
"Person 5: waited at 5 lights."

For the LOW treatment, the screen instead reports 0, 3, 0, 0, and 0 lights waited. The difference in the treatments is thus having four prior participants wait at 5 lights in HIGH or having four prior participants wait at 0 lights in LOW.

The identifying assumption is that the subjects know the rule and that the effect of the treatment is merely to shift beliefs about others' rule following. Remember that the subjects in both treatments are explicitly told that the rule is to wait at red lights, thus creating common knowledge of the rule. Hence, the effect of the treatment should be solely on beliefs about the rate of others' rule following and not awareness of the rule itself.

5.2. Study 3 predictions

The treatments provide noisy signals of what Study 3 subjects should expect in their experiment session. Thus, we expect that Study 3 subjects will update their prior beliefs about others' rule following based on this information. Assuming no aggregate difference in the set of students across the treatments, we expect the beliefs about others' lights waited in the HIGH treatment to be higher than those in the LOW treatment. In the utility function from (2) and (3), we would then have $\beta_i^{HIGH} > \beta_i^{LOW}$ for a subject i assigned to either of those two treatments, and a subject in the HIGH treatment should be more likely than a subject in the LOW treatment to have a high enough belief to follow the rule.

Notice that this last prediction holds for both of the channels discussed in the introduction. By the first channel, an increased (decreased) belief about the rule following of others could increase (decrease) rule following directly as the salience for rule following increases (decreases). By the second channel, believing that enough of the others are following the rule may be necessary for a behavior to be considered a rule at all. An increase in the belief about others' rule followingness can result in the subject treating the rule as being in effect, while a decrease in belief can result in the subject treating the rule as not being in effect. According to both channels, a positive shock to beliefs should increase rule following, while a negative shock should decrease it.

We thus have the following predictions for Study 3.

Hypothesis 3:

(a) The reported belief about others' lights waited will be higher in the HIGH treatment than in the LOW treatment.

(b) The average number of lights waited will be higher in the HIGH treatment than in the LOW treatment.

[16] We note that this method does not involve deception as the selection into the HIGH or LOW treatment is random. Thus, the set of prior participants that is chosen for the subject to learn about is random, and we are not lying to the subjects. A similar experimental design is used in Charness et al. (2019). Note that we did not tell the subjects that the five individuals subjects are selected i.i.d. from a fixed distribution, which would have been untruthful.

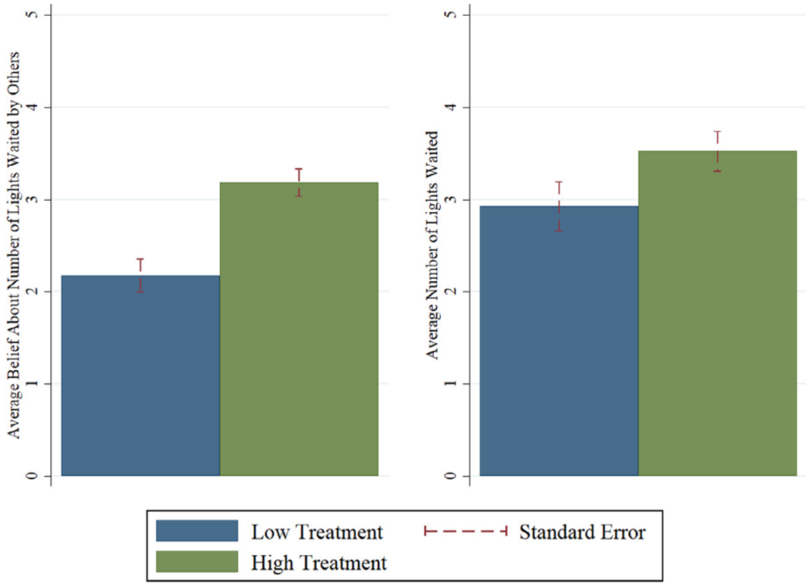

Fig. 5. Study 3- Lights Waited and Beliefs about Lights Waited, Low and High Treatments.

Table 5
Study 3- Predicting Beliefs and Rule-following by High and Low Treatments.

	(1) Belief Other's Number of Lights Waited	(2) Number of Lights Waited	(3) Number of Lights Waited
High Treatment	0.99***	0.59*	-0.35
	(0.23)	(0.34)	(0.32)
Belief Other's Number of Lights Waited			0.94***
			(0.11)
Female	0.38	0.31	-0.05
	(0.26)	(0.41)	(0.31)
Age	-0.07	-0.09	-0.02
	(0.06)	(0.09)	(0.06)
CRT	0.13	-0.01	-0.13
	(0.10)	(0.15)	(0.12)
Intercept	3.19***	4.57**	1.56
	(1.21)	(1.98)	(1.36)
N	139	139	139
R^2	0.155	0.036	0.417

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.3. Study 3 results

Figure 5 plots the average reported belief about others' lights waited by treatment. Subjects reported a significantly higher belief in the HIGH treatment compared to the LOW treatment (Fisher-Pitman two-sided permutation test, $p < 0.001$).[17] The average reported belief in the HIGH treatment is 1 light higher than the average reported belief in the LOW treatment, as seen in regression 1 of Table 5. This difference is highly significant, statistically precise, and qualitatively matches our prediction in Hypothesis 3(a). Also seen in Fig. 5 is the average lights waited by treatment. As predicted in Hypothesis 3(b), subjects waited at more lights in the HIGH treatment than in the LOW treatment (Fisher-Pitman two-sided permutation test, $p = 0.091$). The results from regression 2 of Table 5 show that the HIGH subjects waited at 0.59 more lights on average than the LOW subjects, a statistically significant difference at the 10% level. These results provide our best

[17] For non-parametric tests, we report two-sided Fisher-Pitman permutation tests which are more powerful and superior to the Wilcoxon rank-sum tests if observed values are interval scale (Siegel and Castellan, 1988; Kaiser, 2007). Note that due to the sample size, *p*-values are calculated with Monte Carlo simulations using 200,000 runs (Kaiser, 2007).

causal evidence that rule following is belief conditional. Subjects in the HIGH treatment believe in a higher proportion of others following the rule and therefore follow the rule at higher rates.

6. Study 4: high-low dictator game

6.1. Study 4 procedures

Study 4 replicates Study 1 with one change. At the beginning of the Dictator Game, the subjects read on the screen: “Student participants in a prior experiment completed the same exact task. Five of these prior participants have been randomly selected, and the amount they donated to charity is reported below.” We selected ten subjects from the prior study, put them into two groups – a “high” group and a “low” group. The subject in Study 4 is randomly assigned which group is used for her reference, the first group being the HIGH treatment ($n = 39$), the second group being the LOW treatment ($n = 32$).[18]

For the HIGH treatment, the screen then reports:

“Person 1: donated 5 to charity.”
“Person 2: donated 3 to charity.”
“Person 3: donated 5 to charity.”
“Person 4: donated 5 to charity.”
“Person 5: donated 5 to charity.”

For the LOW treatment, the screen instead reports 0, 3, 0, 0, and 0 as the donations. The difference in the treatments is thus having four prior participants donate 5 in HIGH or have four prior participants donate 0 in LOW.

Note an important difference between Study 3 and Study 4: the rule is explicitly given to the subjects in the RF Task but no behavior is identified as correct in the Dictator Game. We did not explicitly state an injunctive norm in the Dictator Game for multiple reasons. First, KV did not explicitly state an injunctive norm for the decision tasks in their original study. Indeed, the standard way of conducting the Dictator Game does not include any statement about what is an appropriate amount of giving. Second, we want to see if the effect of the treatment found in Study 3 is comparable to that in Study 4 even without stating an injunctive norm. Although there may be two effects of the treatment in Study 4 that correspond to the same two channels mentioned earlier, the overall effect of the treatment on behavior should be the same.

6.2. Study 4 predictions

The treatments serve as a noisy signal of what Study 4 subjects should expect in their experiment session. Thus, we expect that Study 4 subjects will update their prior beliefs about others’ donations based on this signal. Assuming no aggregate difference in the set of students across the treatments, we expect the reported beliefs in the HIGH treatment to be higher than those in the LOW treatment. We should, consequently, also observe a higher level of donations in the HIGH treatment than in the LOW treatment.

Hypothesis 4:

(a) The reported belief about others’ donations will be higher in the HIGH treatment than in the LOW treatment.

(b) The average donations will be higher in the HIGH treatment than in the LOW treatment.

6.3. Study 4 results

The evidence reported in Fig. 6 matches the predictions in Hypothesis 4. Subjects’ reported beliefs about others’ donations are higher in the HIGH treatment compared to the LOW treatment (Fisher-Pitman two-sided permutation test, $p < 0.001$). Similarly, donation levels are higher in the HIGH treatment than in the LOW treatment (Fisher-Pitman two-sided permutation test, $p = 0.003$). Regression results in Table 6 show that subjects in the HIGH treatment expected others to donate $2.52 more on average than subjects in the LOW treatment expected to be donated, and subjects in the HIGH treatment donated $1.62 more than those in the LOW treatment. Exogenous shocks to beliefs about others’ giving thus led to different rates of giving across the two treatments.

As stated earlier, it is possible that the treating of beliefs in Study 4 affects not just the belief about others’ norm-conformity but also affects the belief about what constitutes the injunctive norm. While our experiment cannot rule out the possibility of both channels operating simultaneously, having the simultaneous presence of both channels does match, in spirit, the argument of our paper. Beliefs about others’ norm conformity influence one’s own norm conformity in a causal way. Future work is needed to separate these two channels experimentally.

[18] We note again that this method does not involve deception as the selection into the HIGH or LOW treatment is random. Thus, the set of prior participants that is chosen for the subject to learn about is random.

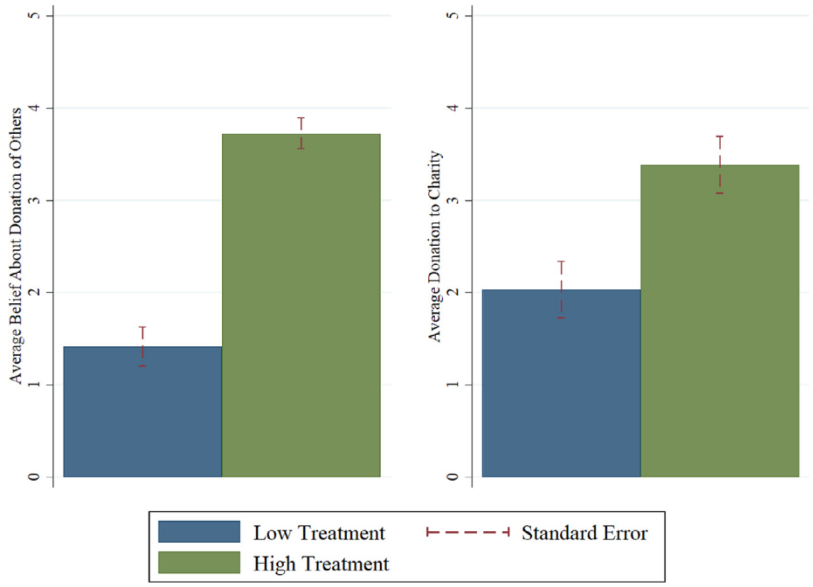

Fig. 6. Study 4- Beliefs about Donations and Donations to Charity, Low and High Treatment.

Table 6
Study 4- Predicting Beliefs and Donations by High and Low Treatments.

	(1)	(2)	(3)
	Belief Other's Donation	Donation	Donation
High Treatment	2.52***	1.62***	-0.13
	(0.25)	(0.44)	(0.85)
Belief Other's Donation			0.69***
			(0.25)
Female	-0.36	-0.12	0.13
	(0.29)	(0.53)	(0.55)
Age	0.17**	0.20*	0.08
	(0.07)	(0.11)	(0.11)
CRT	-0.28*	-0.30	-0.11
	(0.15)	(0.31)	(0.29)
Intercept	-1.74	-1.93	-0.73
	(1.34)	(2.15)	(2.10)
N	69	69	69
R^2	0.571	0.182	0.327

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

7. Additional findings

We here report several additional findings that add further insight into our understanding of the role of beliefs in rule-following behavior.

7.1. Causality from beliefs to behavior

Studies 2 and 3 provide causal evidence that changes in beliefs about others' rule following affects, on average, one's own rule following, while Study 4 provides causal evidence that beliefs about others' giving affects one' own giving. We noted earlier that the presence of an experimenter-demand effect may reduce the plausibility of this interpretation for Study 2, but the evidence from Studies 3 and 4 are less disputable. Indeed, further analysis of Study 3 provided additional evidence of the causal link.

Regression (3) in Table 5 shows that adding the reported belief to regression (2) leads to an insignificant coefficient on the treatment variable in Study 3, which implies that the effect of the treatment in Study 3 appears to operate through beliefs. This finding adds further credibility to our interpretation that the treatment design was successful in creating an exogenous shock to subjects' beliefs about others' rule following. As mentioned earlier, this exogenous shock can work to reduce the negative effect of violating the norm, or it can lead the decision maker into believing that the prescribed behavior

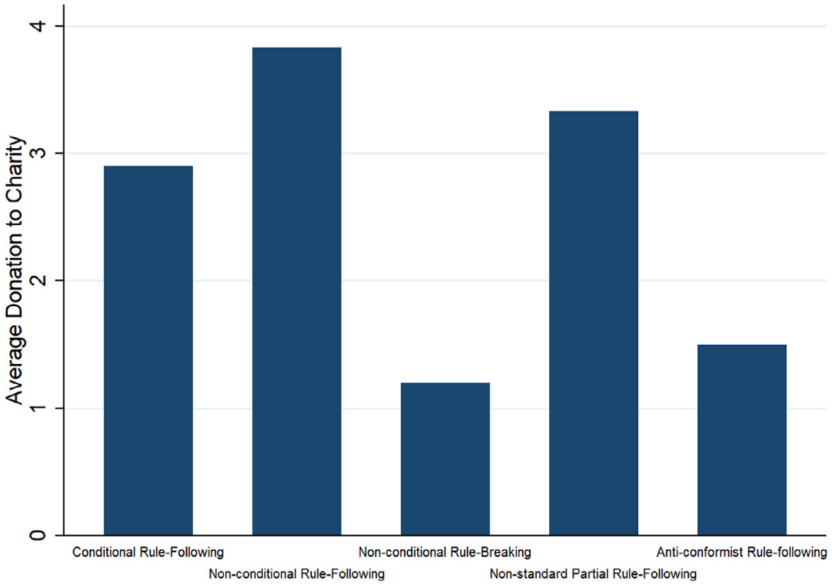

Fig. 7. Study 2- Donations to Charity by Subject Choice Categories in Rule-following Task.

Table 7
Study 2- Subject Choice Categories.

	Number	Percentage
Conditional Rule-following	51	76
Non-conditional Rule-following (Partial or full)	6	9
Non-conditional Rule-breaking (Never waited)	5	8
Non-standard Partial Rule-following	3	4
Anti-conformist Rule-following	2	3
Total	67	1

does not qualify as a norm. This finding also adds credibility to the overall message of our paper that beliefs about others' rule following plays a causal role in rule following. Regression (3) in Table 6 tells a similar story for norm conformity in the form of giving in the Dictator Game.

7.2. Heterogeneity in rule-following

A closer look at the individual strategies submitted by subjects in Study 2 reveals that the majority of subjects (76%) chose a strategy in which lights waited increased as others' lights waited increased. See Fig. 7 and Table 7. The typical subject thus manifests the conditional rule following akin to that in implied by the utility function in Eqs. (2) and (3). However, 16% of the subjects were non-conditional in that they waited at exactly the same number of lights no matter the others' lights waited. Although a utility functions with beliefs such as Eqs. (2) and (3) captures the conditional behavior of the majority of the population, it does not accurately represent all subjects. The Eq. (1) utility function, which does not condition behavior on beliefs, may better represent the non-conditional subjects.

That behavior may condition on others' behavior in some settings but not in others has been a matter of some debate. Bicchieri (2006), for example, distinguishes between social norms for which behavior is conditional and moral norms for which behavior is not, but this distinction has been criticized (Dubreuil and Gregoire, 2013). Our results indicate that the distinction between conditional and non-conditional behavior is empirically relevant when considering individual differences in our setting but that conditional rule-following is the best way to understand the behavior in the aggregate.

7.3. Rule-following and pro-social behavior across settings

One of the key findings from KV is that the propensity to follow rules predicts pro-sociality, and we here add to our understanding of this relationship. If we are correct that rule following and norm conformity are both belief conditional, then we should be able to break that correlation between rule following in the RF Task and norm conformity in the Dictator Game by manipulating beliefs. That is, we should be able to get subjects to follow rules in the RF Task but not give in the Dictator Game or vice versa by appropriately manipulating beliefs.

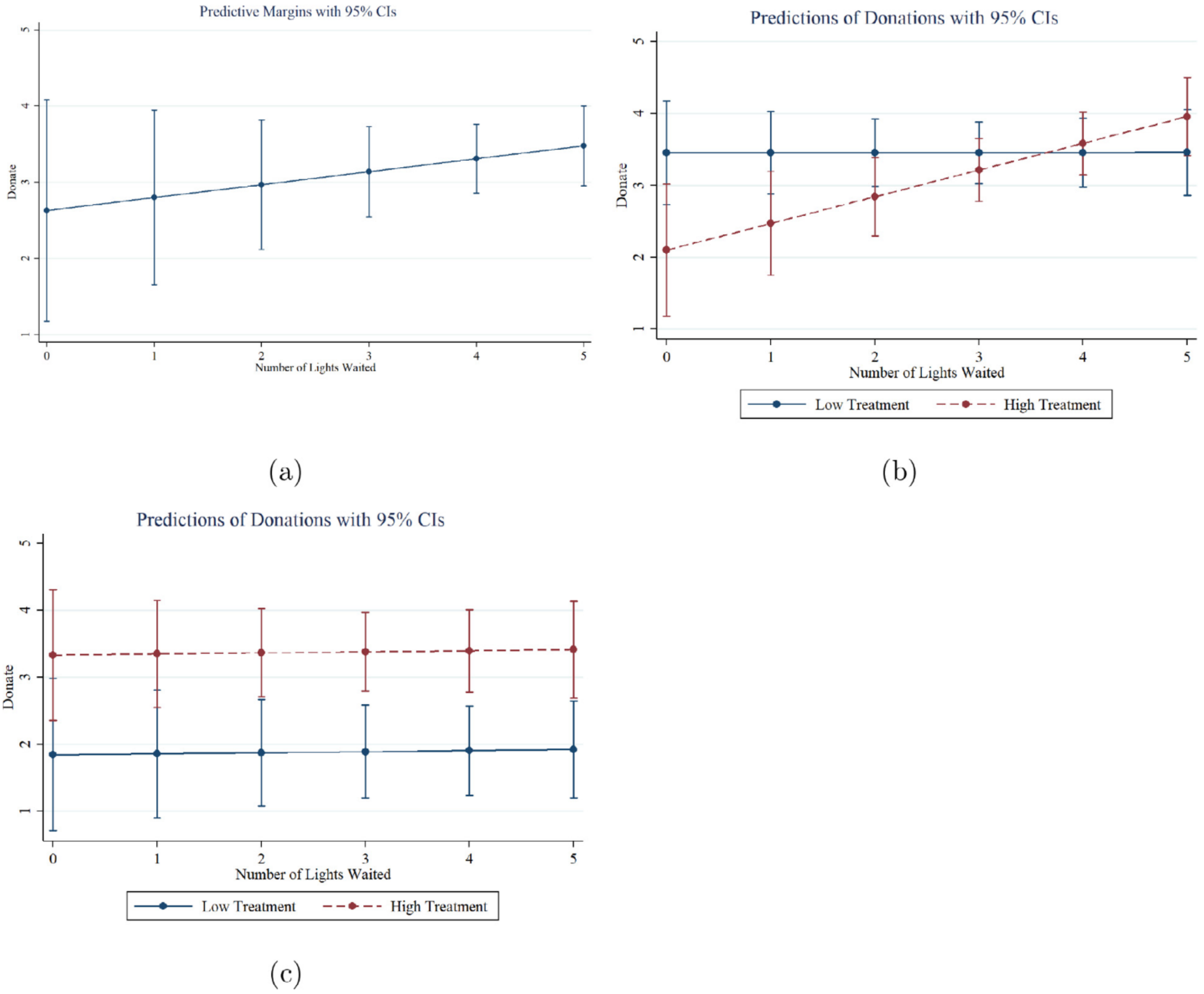

Fig. 8. (a) Study 1: Predicted Donations by Number of Lights Waited. (b) Study 3: Predicted Donations by Number of Lights Waited and by High and Low Treatment. (c) Study 4: Predicted Donations by Number of Lights Waited and by High and Low Treatment.

Figure 8 (a) plots the predicted donations by the Number of Lights Waited in Study 1, and it shows a positive but insignificant linear relationship between rule-following and donations. Figure 8(b) shows how the break occurred in our Study 3. In the Study 3 HIGH treatment, subjects who waited at fewer lights donated less on average than subjects who waited at several lights, which matches the KV finding. However, this correlation disappears in the Study 3 LOW treatment, i.e., the subjects that waited at fewer lights donated just as much as the subjects that waited at many lights (see Table 8 column 2). Regression analysis for Study 1 showed that the correlation between rule-following and donations was strongest when comparing those who waited at zero lights to those who waited at one or more lights.[19] Figure 9 demonstrates the break in correlation when comparing those that never waited with those who waited at least once. We see the correlation in Fig. 9(a) for Study 1 and in Fig. 9(c) for the HIGH treatment of Study 3, but not in Fig. 9(b) for the LOW treatment in Study 3. Study 3 LOW subjects who chose to never follow the rules donated similar amounts to those who waited at one or more lights (see Table 9 column 1).

One interpretation is that the HIGH treatment had a smaller effect on beliefs relative to the Study 1 baseline than did the LOW treatment, the result being that the LOW-treatment subjects who believed there would be very little rule following still believed there would be norm conformity in the Dictator Game. Put differently, there was little spillover in beliefs from rule following in the RF Task to norm conformity in the Dictator Game. This interpretation fits our argument that the link between rule following and norm compliance found by KV depends on the presence of a correlation in beliefs about others'

[19] Regression analysis using all the categorical variables for rule-following for Study 3 LOW and HIGH Treatments can be found in the Supplemental Appendix (See Tables A1 and A2).

Table 8
Studies 3 and 4: Predicting Donation by Number of Lights Waited by High and Low Treatments.

	(1)	(2)	(3)	(4)
	Study 3		Study 4	
	Donate	Donate	Donate	Donate
Number of Lights Waited	0.14*	-0.01	0.02	0.00
	(0.08)	(0.10)	(0.12)	(0.18)
High Treatment	-0.15	-1.39**	1.63***	1.56*
	(0.31)	(0.64)	(0.46)	(0.84)
Number of Lights Waited x High Treatment		0.38**		0.02
		(0.16)		(0.23)
Female	0.49	0.50	-0.12	-0.12
	(0.35)	(0.38)	(0.50)	(0.55)
Age	-0.04	-0.05	0.20	0.20*
	(0.08)	(0.08)	(0.13)	(0.11)
CRT	0.03	0.05	-0.30	-0.30
	(0.14)	(0.14)	(0.30)	(0.32)
Intercept	3.49**	4.00**	-2.00	-1.94
	(1.67)	(1.75)	(2.68)	(2.18)
N	139	139	69	69
R^2	0.046	0.087	0.183	0.183

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

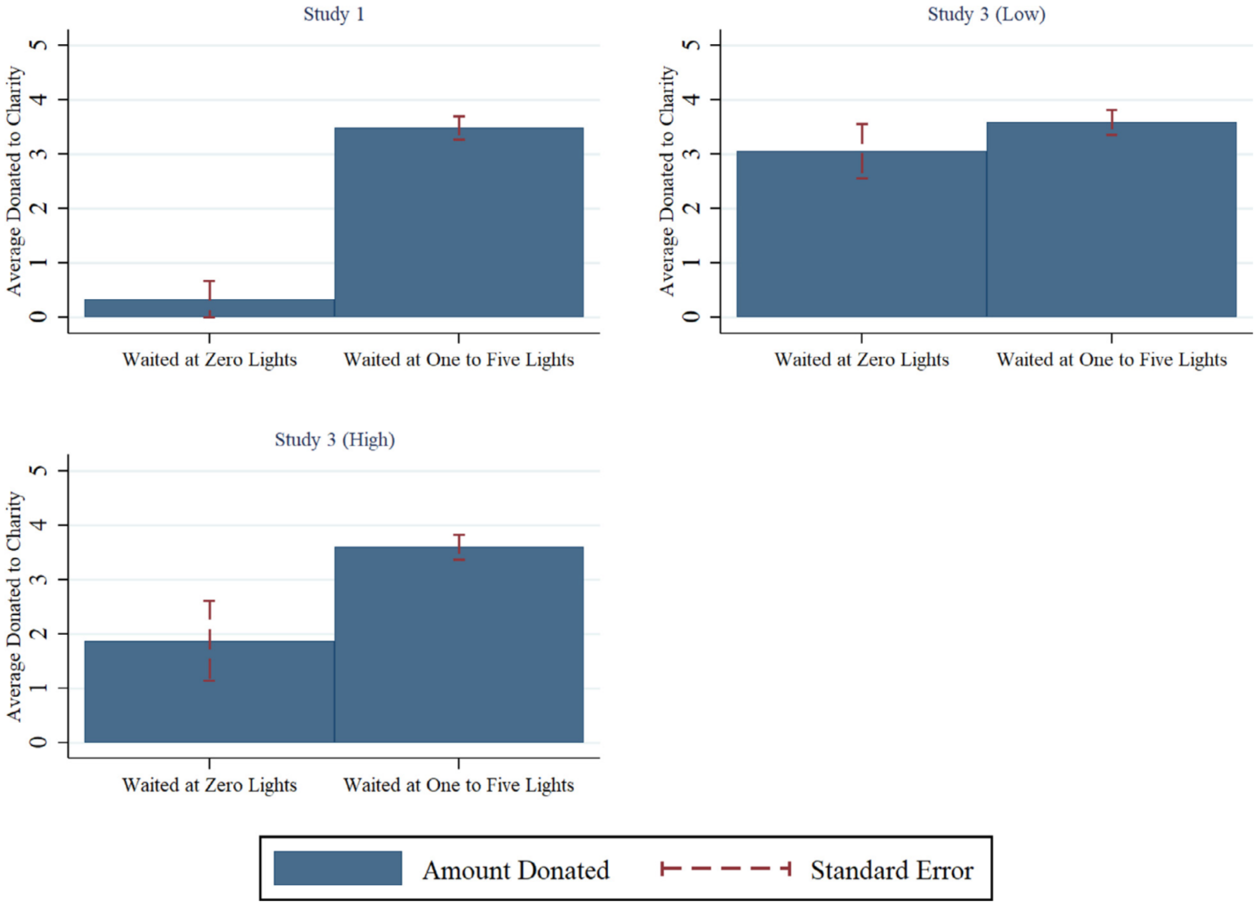

Fig. 9. Study 1 and Study 3: Donations by Category of Number of Lights Waited.

Table 9
Studies 3 and 4: Predicting Donation by Category of Number of Lights Waited.

	(1)	(2)	(3)	(4)
	Study 3		Study 4	
	Low Donate	High Donate	Low Donate	High Donate
Waited at One to Five Lights	0.31	1.82**	-0.02	0.02
	(0.55)	(0.78)	(0.77)	(0.45)
Female	0.96*	0.10	-0.10	-0.37
	(0.56)	(0.58)	(0.91)	(0.66)
Age	-0.00	-0.05	0.26**	0.11
	(0.12)	(0.11)	(0.11)	(0.23)
CRT	0.04	0.18	-0.73	0.12
	(0.21)	(0.18)	(0.45)	(0.0.43)
Intercept	2.52	2.48	-1.12	-0.22
	(2.63)	(2.70)	(2.28)	(4.73)
N	68	71	39	30
R^2	0.075	0.097	0.11	0.04

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

behavior across the settings. Those that believe others will wait at many lights are also likely to believe that others will donate higher amounts to charity. It is not a fixed personality trait alone that holds across settings but rather a correlation in beliefs across settings combined with a rule-following propensity.

The break is even more striking in Study 4. As seen in Fig. 8(c), there is no correlation between donating and lights waited for either treatment in Study 4, indicating a clean break between rule following and norm conformity across the two settings (see Table 8 column 4). Unlike in Study 1, those that always followed the rule and those that never followed the rule within the HIGH treatment donated identical amounts, and the same holds for the LOW treatment (see Table 9 columns 3 and 4).[20] This finding matches a claim that rule following and norm conformity are both belief conditional. It does not match a claim that rule following and norm conformity depend only on individual, fixed dispositions.

8. Conclusion

We argued and presented experimental evidence that rule-following behavior, just like norm-conforming behavior, is belief conditional. In particular, adherence to social norms depends on both a rule-following propensity as well as a belief that others will follow the rules. This evidence supports a richer understanding of human decision making that is consistent with existing literatures within and outside of economics that incorporates social norms into utility representations.

The most immediate lesson from our four studies is that social-norm representations of preferences should include beliefs about others' behavior. Although some of our subjects choose strategies that are not conditional on beliefs, the large majority exhibit belief-conditional behavior. Belief conditionality is thus the best way to understand the aggregate behavior of the subjects across the studies.

Another lesson learned is that the behavior in KV's RF Task reflects both beliefs about the compliance of others and a fixed preference to follow norms. This finding is not a criticism of the RF Task. On the contrary, that we could so easily adapt the RF Task provides another demonstration of its value as a research tool in studying human pro-sociality.

There still remain many important avenues for future research. One potentially fruitful line of inquiry asks how it is that subjects come to understand what is the relevant injunctive norm in a setting when no norm is explicitly stated. Another explores heterogeneity in rule following and norm compliance across settings and subjects. Do subjects that manifest belief-conditional behavior in one setting always manifest belief-conditional behavior in other settings? And are subjects that exhibit non-conditional behavior more likely to be non-conditional in other settings? Finding answers to these questions will provide additional insights into our understanding of pro-social behavior and enable researchers to better represent human preferences in their models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

[20] Regression analysis using all the categorical variables for rule-following for Study 4 LOW and HIGH Treatments can be found in the Supplemental Appendix (See Tables A3 and A4).

Appendix A. Linear utility analysis

We here show that for each i there will exist a $\beta_i^* \geq 0$ such that i will follow the norm if and only if $\beta_i \geq \max\left\{\beta_i^*, \widehat{\beta}\right\}$. This result provides the main logic behind the predicted correlation between beliefs and rule following and between beliefs and norm conformity in the experimental studies. For ease of analysis, we consider a following linear version of the utility function from Eqs. (2) and (3) as this allows for an explicit derivation of β_i^*. If player i has a non-linear monetary payoff function or a non-linear disutility for violating the norm, then a similar result obtains except β_i^* will be derived implicitly.

Consider the following linear utility function in which player i's monetary payoff depends only on their own decision:

$$u_i(s_i, s_{-i}) = M - as_i - k_i\beta_i g(|s_i - \widehat{s}|, \beta_i), \tag{4}$$

$$g(|s_i - \widehat{s}|, \beta_i) = \begin{cases} |s_i - \widehat{s}|, & \text{if } \beta_i > \widehat{\beta}, \\ 0, & \text{if } \beta_i \leq \widehat{\beta}, \end{cases} \tag{5}$$

with $a > 0$ and $k_i > 0$. This linear version satisfies the conditions for the utility function in Eqs. (2) and (3), i.e., the norm (or rule) salience is strictly increasing in β_i and in the distance between s_i and the social norm $\widehat{s}$.

To further simplify, we will restrict the strategy space such that $s_i \in \left[0, \frac{M}{a}\right]$ to ensure that i's monetary payoff is never negative. The social norm has the same domain of $\left[0, \frac{M}{a}\right]$.

Moreover, our analysis here will apply to both the RF Task and the Dictator Game. In both scenarios, the monetary payoff depends on the subject's own choice but not on other subjects' actions. The disutility from deviating from the norm (or rule) does, however, depend on the belief about others' behavior.

Player i's optimal behavior is best illustrated by distinguishing three cases.

Case I: Low belief $\beta_i \leq \widehat{\beta}$. With low conformity, i's utility function becomes

$$u_i(s_i, s_{-i}) = M - as_i.$$

With $a > 0$, utility is maximized by choosing the smallest s_i, i.e., $s_i = 0$. Player i does not conform because i believes that too few others are conforming.

Case II. High belief $\beta_i > \widehat{\beta}$**, low salience** $k_i < \frac{a}{\beta_i}$**.** First consider $s_i \leq \widehat{s}$. The utility function is

$$u_i(s_i, s_{-i}) = M - as_i - k_i\beta_i(\widehat{s} - s_i).$$

The derivative with respect to s_i is

$$\frac{\partial u_i(s_i, s_{-i})}{\partial s_i} = -a + k_i\beta_i.$$

With $k_i < \frac{a}{\beta_i}$, this derivative is strictly negative, so i's highest utility over $[0, \widehat{s}]$ is obtained with $s_i = 0$.

Now consider $s_i \geq \widehat{s}$. The utility function and first derivative are now

$$u_i(s_i, s_{-i}) = M - as_i - k_i\beta_i(s_i - \widehat{s}),$$
$$\frac{\partial u_i(s_i, s_{-i})}{\partial s_i} = -a - k_i\beta_i,$$

and this derivative is always negative. Thus, over the range $\left[\widehat{s}, \frac{M}{a}\right]$, the utility is maximized by setting $s_i = \widehat{s}$.

Combining our analysis of $[0, \widehat{s}]$ and $\left[\widehat{s}, \frac{M}{a}\right]$, we see that i maximizes utility in this case by setting $s_i = 0$. Player i's norm salience is too low to motivate norm compliance even though many others are believed to be conforming.

Case III. High belief $\beta_i > \widehat{\beta}$ **, high salience** $k_i > \frac{a}{\beta_i}$ **.** First consider $s_i \leq \widehat{s}$. The utility function and first derivative are

$$u_i(s_i, s_{-i}) = M - as_i - k_i\beta_i(\widehat{s} - s_i),$$
$$\frac{\partial u_i(s_i, s_{-i})}{\partial s_i} = -a + k_i\beta_i.$$

With $k_i > \frac{a}{\beta_i}$, this derivative is strictly positive, so i's highest utility over $[0, \widehat{s}]$ is obtained with $s_i = \widehat{s}$.

The analysis for $s_i \geq \widehat{s}$ is identical to that in Case II, with utility over $\left[\widehat{s}, \frac{M}{a}\right]$ maximized by setting $s_i = \widehat{s}$.

Combining our analysis of $[0, \widehat{s}]$ and $\left[\widehat{s}, \frac{M}{a}\right]$, we see that i maximizes utility in this case by setting $s_i = \widehat{s}$. Player i's norm salience is high enough to motivate norm compliance.

After consideration of all three cases, we now define $\beta_i^* = \frac{a}{k_i}$ and observe that if $\beta_i \geq \max\left\{\beta_i^*, \widehat{\beta}\right\}$, then i will be in Case III with best response $s_i = \widehat{s}$. Moreover, if $\beta_i < \max\left\{\beta_i^*, \widehat{\beta}\right\}$, then i's best response is not $s_i = \widehat{s}$. Thus, $\beta_i \geq \max\left\{\beta_i^*, \widehat{\beta}\right\}$ is necessary and sufficient for i to conform to the norm. This establishes the claim.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2021.09.030.

References

Andreoni, J., Bernheim, D., 2009. Social image and the 50-50 norm: a theoretical and experimental analysis of audience effects. Econometrica 77, 1607–1636.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., Plumb, I., 2001. The “reading the mind in the eyes” test revised version: a study with normal adults, and adults with Asperger syndrome or high-functiong austism. J. Child Psychol. Psychiatry 42, 241–251.

Benabou, R., Tirole, J., 2006. Incentives and prosocial behavior. Am. Econ. Rev. 96, 1652–1678.

Bicchieri, C., 2006. The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press.

Bicchieri, C., 2008. The fragility of fairness: an experimental investigation on the conditional status of pro-social norms. Philos. Issues 18, 229–248.

Bicchieri, C., Xiao, E., 2009. Do the right thing: but only if others do so. J. Behav. Decis. Mak. 22, 191–208.

Blanco, M., Engelmann, D., Koch, A.K., Normann, H.T., 2010. Belief elicitation in experiments: is there a hedging problem? Exp. Econ. 13, 412–438.

Bowles, S., Gintis, H., 2011. A Cooperative Species: Human Reciprocity and Its Evolution. Princeton University Press.

Boyd, R., Richerson, P.J., 1988. Culture and the Evolutionary Process. Chicago: University of Chicago Press.

Brandts, J., Charness, G., 2011. The strategy versus the direct-response method: a first survey of experimental comparisons. Exp. Econ. 14, 375–398.

Brunner, M., Ostermaier, A., 2019. Peer influence on managerial honesty: the role of transparency and expectations. J. Bus. Ethics 154, 127–145.

Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74, 1579–1601.

Charness, G., Naef, M., Sontuoso, A., 2019. Opportunistic conformism. J. Econ. Theory 180, 100–134.

Cialdini, R.B., Reno, R.R., Kallgren, C.A., 1990. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places.

Desmet, P., Engel, C., 2017. People are Conditional Rule Followers. MPI Collective Goods Preprint, No. 2017/09 doi:10.2139/ssrn.2969799.

Diekmann, A., Przepiorka, W., Rauhut, H., 2015. Lifting the veil of ignorance: an experiment on the contagiousness of norm violations. Rationality Soc. 27, 309–333.

Dimant, E., 2017. On Peer Effects: Contagion of Pro- and Anti-Social Behavior and the Role of Social Cohesion. Discussion Paper No. 2017-06. The University of Nottingham, Center for Decision Research and Experimental Economics (CEDEX).

Dreber, A., Ellignsen, T., Johannesson, M., Rand, D.G., 2013. Do people care about social context? Framing effects in dictator games. Exp. Econ. 16, 349–371.

Dubreuil, B., Gregoire, J.-F., 2013. Are moral norms distinct from social norms? A critical assessment of Jon Elster and Cristina Bicchieri. Theory Decis. 75, 137–152.

Elster, J., 1989. Social norms and economic theory. Am. Econ. Rev. 3, 99–117.

Elster, J., 2009. The Oxford Handbook of Analytical Sociology. Oxford University Press, pp. 195–217. Ch. Norms

Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114 (3), 817–868.

Fershtman, C., Gneezy, U., List, J.A., 2012. Equity aversion: social norms and the desire to be ahead. Am. Econ. J. 4, 131–144.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10, 171–178.

Gachter, S., Gerhards, L., Nosenzo, D., 2017. The importance of peers for compliance with norms of fair sharing. Eur. Econ. Rev. 97, 72–86.

Gachter, S., Nosenzo, D., Sefton, M., 2013. Peer effects in pro-social behavior: social norms or social preferences? J. Eur. Econ. Assoc. 11, 548–573.

Gino, F., Ayal, S., Ariely, D., 2009. Contagion and differentiation in unethical behavior. Psychol Sci 20, 393–398.

Heath, J., 2008. Following the Rules: Practical Reasoning and Deontic Constraint. Oxford University Press.

Kaiser, J., 2007. An exact and a monte carlo proposal to the Fisher-Pitman permutation tests for paired replicates and for independent samples. Stata J 402–412.

Kamei, K., 2014. Conditional punishment. Econ. Lett. 124, 199–202.

Kessler, J.B., Leider, S., 2012. Norms and contracting. Manage. Sci. 58, 62–77.

Kimbrough, E.O., Vostroknutov, A., 2016. Norms make preferences social. J. Eur. Econ. Assoc. 14 (3), 608–638.

Kroher, M., Wolbring, T., 2015. Social control, social learning, and cheating: evidence from lab and online experiments on dishonesty. Soc. Sci. Res. 53, 311–324.

Krupka, E., Weber, R.A., 2009. The focusing and informational effects of norms on pro-social behavior. J. Econ. Psychol. 30, 307–320.

Krupka, E.L., Weber, R.A., 2013. Identifying social norms using coordination games: why does dictator game sharing vary? J. Eur. Econ. Assoc. 11, 495–524.

Lahno, A., Serra-Garcia, M., 2015. Peer effects in risk taking: envy or conformity? J. Risk Uncertain. 50, 73–95.

List, J.A., 2007. On the interpretation of giving in dictator games. J. Polit. Economy 115, 482–493.

Lopez-Perez, R., 2008. Aversion to norm-breaking: a model. Games Econ. Behav. 64, 237–267.

Richerson, P.J., Boyd, R., 2008. Not By Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press.

Ridinger, G., 2018. Ownership, punishment, and norms in a real-effort bargaining experiment. J. Econ. Behav. Organ. 155, 382–402. doi:10.1016/j.jebo.2018.09.008.

Ridinger, G., McBride, M., 2020. Reciprocity in games with unknown types. In: Capra, M., Croson, R., Rigdon, M., Rosenblatt, T. (Eds.), Handbook of Experimental Game Theory. Edward Elgar Publishing, pp. 271–288.

Schram, A., Charness, G., 2015. Inducing social norms in laboratory allocation choices. Manage. Sci. 61 (7), 1531–1546.

Siegel, S., Castellan, N.J., 1988. Nonparametric Statistics for the Behavioral Sciences, second ed. New York: McGraw Hill.

Thoni, C., Gachter, S., 2015. Peer effects and social preferences in voluntary cooperation: a theoretical and experimental analysis. J. Econ. Psychol. 48, 72–88.