From good institutions to good norms: Top-down incentives to cooperate foster prosociality but not norm enforcement

Michael N. Stagnaro[1*†], Antonio A. Arechar[1†], David G. Rand[1,2,3]
[1]Department of Psychology, [2]Department of Economics, [3]School of Management, Yale University
*Corresponding author: michael.stagnaro@yale.edu [†]These authors contributed equally to this work.

What makes people willing to pay costs to help others, and to punish others' selfishness? Why does the extent of such behaviors vary markedly across cultures? To shed light on these questions, we explore the role of formal institutions in shaping individuals' prosociality and punishment. In Study 1 (N=707), we found that the quality of institutions enforcing cooperativeness (police and courts) that American participants reported being exposed to in daily life was positively associated with Dictator Game (DG) giving, but had no significant relationship with punishment in a Third-Party Punishment Game (TPPG). In Study 1R (N=1,705), we replicated the positive relationship between reported institutional quality and DG giving. In Study 2 (N=516), we experimentally manipulated institutional quality in a repeated Public Goods Game with a centralized punishment institution. Consistent with Study 1's correlational results, we found that centralized punishment led to significantly more prosociality in a subsequent DG compared to a no-punishment control, but had no significant direct effect on subsequent TPPG punishment (only an indirect effect via increased DG giving). Thus we present convergent evidence that the quality of institutions one is exposed to "spills over" to subsequent prosociality but not punishment. These findings support a theory of social heuristics, suggest boundary conditions on spillover effects of cooperation, and demonstrate the power of effective institutions for instilling habits of virtue and creating cultures of cooperation.

1. Introduction

The willingness to do what is right, even when it is personally costly, is a central part of what makes society function. In particular, people regularly bear costs in order to create benefits for others – that is, they cooperate. For decades, researchers across the social and natural sciences have investigated what makes individuals choose to engage in cooperation despite the personal costs involved (Axelrod & Hamilton, 1981; Batson, Duncan, Ackerman, Buckley, & Birch, 1981; R. Campbell & Sowden, 1985; Chakroff & Young, 2014; Chudek & Henrich, 2011; Cushman & Macindoe, 2009; Galinsky & Schweitzer, 2015; Hamlin, Wynn, & Bloom, 2007; Kiyonari, Tanida, & Yamagishi, 2000; Kraft-Todd, Yoeli, Bhanot, & Rand, 2015; Paluck, Shepherd, & Aronow, 2016; Raihani, Thornton, & Bshary, 2012; Rand & Nowak, 2013; Tooby & Cosmides, 1990; Van Lange, De Bruin, Otten, & Joireman, 1997).

One class of answers comes in the form of mechanisms that resolve the dilemma of cooperation by creating future benefits for cooperation, and/or future punishments for failing to cooperate (e.g. repeated interactions, reputation systems, ostracism, and sanctioning institutions; for a review, see Rand and Nowak (2013)). The incentives provided by these mechanisms make it so that even self-interested individuals may choose to cooperate (e.g. Dreber, Fudenberg, and Rand (2014)), and such "strategic" cooperation likely accounts for much of the everyday cooperative behaviors that typify human societies.

Yet it is also clear that people often engage in "pure" cooperation, cooperating in situations where it is truly not in one's material self-interest to do so – for example, in the context of 1-shot anonymous interactions. Importantly, a robustly observed feature of such pure cooperation is that, despite playing a key role in promoting the collective good, the extent to which people cooperate with strangers varies markedly across cultures (Blake et al., 2015; Cappelen, Moene, Sørensen, & Tungodden, 2013; Gächter, Herrmann, & Thöni, 2010; Henrich et al., 2005; Henrich et al., 2010; Sapienza, Zingales, & Guiso, 2006; Yamagishi, Hashimoto, & Schug, 2008).

What makes people willing to engage in pure cooperation in the first place? And what explains large-scale variation in that willingness? Here, we seek to shed new light on these questions using insights regarding the *cognitive* underpinnings of cooperation. We argue that formal institutions which increase accountability and enforcement create top-down incentives to cooperate, leading people to adopt heuristics prescribing cooperation which get applied even in pure cooperation settings beyond the reach of any institution; and that cross-cultural variation in the quality of these formal institutions therefore helps to explain cross-cultural variation in levels of pure cooperation.

Central to this reasoning is the idea that any given cooperation decision is shaped not just by the details of the situation (e.g. presence or absence of incentives to cooperate), but also by what behaviors have worked well in the context of the decider's past experiences in other situations (Bear & Rand, 2016; Chudek & Henrich, 2011; Kiyonari et al., 2000; Rand et al., 2014; Tomasello,

Melis, Tennie, Wyman, & Herrmann, 2012; Tooby & Cosmides, 1990; Van Lange et al., 1997; Yamagishi et al., 2008). In particular, prior experience with situations where cooperation *is* in one's long-run self-interest may lead one to sometimes cooperate even in the context of pure cooperation – the incentives to cooperate that exist in many settings may "spill over" into settings where such incentives are lacking.

We focus here on a particular spillover-based theory that is more explicitly cognitive than most others: the Social Heuristics Hypothesis (SHH) (Bear & Rand, 2016; Rand et al., 2014). The SHH applies the dual-process model of decision-making (Evans, 2008; Kahneman, 2003; Sloman, 1996; Stanovich & West, 2000), where decisions are conceptualized as resulting from the interaction of relatively intuitive versus deliberative cognitive processes, to the consideration of spillover effects. The SHH argues that people develop heuristics for social interaction that are shaped by daily social experiences. These social heuristics provide fast, rule-of-thumb implementations of strategies that are typically successful and therefore become automatized as default responses. It then requires deliberation to override these defaults in atypical situations where the incentives do not match those of daily life. As a result, intuitive responses tend to treat (atypical) pure cooperation settings as if they were (typical) interactions involving future consequences (and therefore incentives to cooperate). The SHH provides an adaptive logic for spillovers: because deliberation is often costly, it can be advantageous to rely on cognitively efficient but inflexible social heuristics, especially when one must make decisions quickly or one is distracted or fatigued (for a formal evolutionary game theoretic demonstration, see Bear and Rand (2016)). Empirical support for the SHH comes, for example, from the observation that experimentally inducing participants to decide more deliberatively (or less intuitively) leads to less cooperation – as shown in a meta-analysis of 51 pure cooperation studies where cognitive processing mode was experimentally manipulated using cognitive load, time constraints, ego depletion, or intuition inductions (Rand, 2016b).

In addition to explaining why people should ever engage in pure cooperation at all, this framework also offers an explanation for cross-cultural variation in pure cooperation: institutional incentives to cooperate, the quality of which varies markedly across cultures. Formal institutions (e.g. legal systems, financial markets, religious codes, or organizational rules) create top-down incentives designed to shape behavior, often punishing wrong-doers, compensating victims, and rewarding those who act to benefit the greater good. Based on this logic, when institutions are strong, there are correspondingly strong incentives to cooperate in typical interactions – and this should lead people living under those institutions to develop cooperative social heuristics. Conversely, when institutions are weak (e.g. corrupt legal or political systems, ineffective policing), it is often not in one's self-interest to cooperate, leading to the development of social heuristics that prescribe selfishness. Therefore, we predict that there should be a causal relationship between institutional quality and pure cooperation in settings beyond the reach of institutional incentives. In other words, the institutions that govern one's life help to shape one's notions of right and wrong.

Cross-cultural studies provide preliminary support for this prediction. For example, Henrich et al. (2010) find a significant positive relationship between a community's level of market integration (an institution which facilitates transactions between strangers) and average giving in the Dictator Game (DG) and a marginally significant positive relationship between adherence to world religions (institutions which prescribe benevolence to strangers) and DG giving. The DG is the most basic measure of pure prosociality, wherein the participant unilaterally splits a sum of money with an anonymous stranger (Forsythe, Horowitz, Savin, & Sefton, 1994). Relatedly, in online studies, Raihani, Mace, and Lamba (2013) find that American participants give more in the DG than Indian participants, and Capraro and Cococcioni (2015) find that Indian participants cooperate less in a Prisoner's Dilemma (PD) compared to the identical PD played by American participants in Capraro, Jordan, and Rand (2014). With respect to punishment, examining different communities within the United States has shown that people in areas lacking strong centralized institutions are more tolerant of certain kinds of moral violations, such as violence in service of protecting one's self and one's property – in these contexts, a "culture of honor" can develop such that acts of aggression and dominance are even encouraged and rewarded with reputational gains (Anderson, 2000; J. K. Campbell, 1966; Cohen, Bowdle, Nisbett, & Schwarz, 1996; Edgerton, 1971). Henrich et al. (2010) find similar variation in the extent to which people choose to punish selfishness, and Herrmann, Thoni, and Gächter (2008) find that a country's strength of rule of law is positively related to the extent to which punishment is selectively targeted at non-cooperators.

Thus, there is correlational evidence suggesting the predicted relationship between the quality of cooperation-enforcing institutions and the extent of cooperative behavior in settings beyond the reach of those institutions. However, these correlational studies do not of course establish the causal arrow between institutions and norms that we suggest. These results are equally consistent with the opposite argument that good norms are required to support strong institutions; or the observed associations could be the result of another unobserved factor that drives both institutional quality and cooperation levels.

Some reason to believe that the predicted causal relationship does indeed exist comes from a recent laboratory study examining the link between peer-based reputational incentives to cooperate and subsequent prosociality (Peysakhovich & Rand, 2016). In the first experimental stage, participants played a series of repeated PD games, with the probability of interacting again with the same partner (and therefore being accountable for your current actions) varying between experimental conditions. Then in the second stage, they played a battery of one-shot anonymous games involving pure cooperation (including a DG) or punishment. The results provide clear evidence of spillover effects: participants randomized into the low accountability condition for the PD games in the first stage were substantially less likely to engage in pure cooperation, or to punish selfishness, in the one-shot anonymous games of the second stage. Further evidence in this vein comes from an experiment where third party observers could pay to intervene and thereby reduce the payoff of defectors and increase the payoff of cooperators (Nakashima, Halali, & Halevy, 2016). Intervention was found to increase cooperation, and when intervention was possible only

in the first half of the game, the increased cooperation carried over into the second intervention-free half of the game.

But does this connection between peer-based reputational incentives to cooperate and subsequent prosociality extend to formal institutions? Although our theory predicts that the answer will be "yes," there are reasons to worry that formal institutions that enforce cooperation will not work in the same way as peer incentives. Specifically, there is evidence that in some settings, extrinsic top-down motivations to behave in a given way undermine (or "crowd-out") one's intrinsic motives to do so (Deci, Koestner, & Ryan, 1999; Frey & Jegen, 2001; Titmuss, 1970). For example, Lepper, Greene, and Nisbett (1973) found that children who were rewarded for drawing in a primary task were significantly less likely to freely choose drawing for an activity at a later time, compared to those who received no expected reward to draw. Relatedly, Gneezy and Rustichini (2000) found that implementing a monetary fine for parents picking their children up late at an Israeli daycare *increased* parents' tardiness - the intrinsic motivation of being a good citizen (not picking your child up late, forcing the staff to stay later) was crowded out by the new transactional frame (the staff are being paid to stay late and watch the kids).

It is not clear, therefore, whether formal institutions that incentivize cooperation will in fact foster prosociality via spillovers, as predicted by the SHH and other spillover theories. Additionally, we examine the relationship between institutions and both one's own prosociality (as measured by the DG), as well as the *enforcement* of prosocial behavior (as measured by a third-party punishment game, TPPG (Fehr & Fischbacher, 2004; Jordan, McAuliffe, & Rand, 2015). Although prosociality and punishment are often discussed as two sides of the same coin, there is empirical evidence suggesting that they are in fact psychologically/motivationally distinct: two recent studies looking at correlation in play across games found no association between cooperativeness and punishment (Peysakhovich, Nowak, & Rand, 2014; Yamagishi et al., 2012), and another showed evidence that punishment can be driven by spiteful (rather than "altruistic") motives (Espín, Brañas-Garza, Herrmann, & Gamella, 2012). That is not to say that cooperation and punishment are totally unrelated – for example, there is also evidence that punishment can be used as a signal of one's cooperativeness (Barclay, 2006; Jordan, Hoffman, Bloom, & Rand, 2016; Raihani & Bshary, 2015). But it is clear that the psychological forces driving one to cooperate differ in important ways from those driving one to punish.

Examining whether institutions have similar effects on prosociality and punishment also helps to illuminate the mechanism underlying spillover effects. If spillovers are driven by social heuristics, as argued by the SHH, we would see increased prosociality due to spillovers, but would have little reason to expect an increase in norm enforcement. It is straightforward to see why the SHH predicts that an institution that incentivizes cooperation should foster a heuristic for paying costs to benefit others (which is then applied in the DG), as that is the precise behavior which is being incentivized by the institution. There is no reason to expect punishment, however, to be affected in the same way – institutional incentives that change the payoff associated with cooperation do not alter the payoffs associated with punishment. The prediction is different,

however, if spillovers are driven by changing understandings of relevant social norms, rather than by heuristics. By this alternate account, seeing other participants cooperate under strong institutions strengthens one's belief that being prosocial is normative in the current setting (and seeing others defect under weak institutions undermines this belief). And therefore, to the extent that people punish behaviors they see as norm violations, we should expect to see more punishment following exposure to stronger institutions. Thus, examining whether incentives to cooperate spill over to influence punishment helps to distinguish between the SHH and norms-based accounts of spillovers.

In the current work, we provide insight into the relationship between cooperation, punishment, and formal institutions that hold citizens accountable and dissuade selfishness with two studies. In Study 1, we provide external validity by correlating a measure of real-world institutional quality with participants' prosociality in the DG and punishment in the TPPG. In Study 2, we demonstrate causality by experimentally varying the quality of a centralized punishment institution imposed on participants playing a repeated cooperation game, and measuring the effects on subsequent DG and TPPG play.

2. Study 1: Correlating institutional quality with prosociality and punishment

2.1. Materials and methods

2.1.1. Participants

We recruited 707 participants via Amazon Mechanical Turk (MTurk; Horton, Rand, and Zeckhauser (2011)), restricting the geographical location of our subjects to the USA. The average age of participants in this sample was 32 years (min 18, max 71), and 48.4% were female. The task lasted between 5 and 10 minutes and participants received a flat payment of $0.40 for participating, plus a variable bonus that on average totaled $0.35 among those who passed the comprehension questions and thus received a bonus (min $0.13, max $0.43). We prevented repeated participation by excluding duplicate worker IDs and IP addresses.

2.1.2. Method

The study consisted of three stages (presented in random order, as described below). The first stage assessed the quality of institutions related to accountability and punishment which participants were exposed to in their daily lives. To do so, we asked them to self-report their confidence in the police and in the courts (each rated using a four-point Likert scale ranging from "A great deal of confidence" to "None at all").[1] The second stage assessed participants'

[1] These two items were taken from a Worlds Values Survey scale with six items total, which also asked about government, political parties, civil services, and the banking industry. Using the full six-item measure gives qualitatively equivalent results (see Appendix A).

prosociality by having them play a single-shot DG with another MTurk worker (MTurker), in which the Dictator unilaterally chose how to divide 30 cents (in 2 cent increments) between herself and the Recipient. The third stage assessed participants' punishment behavior by having them play a single-shot TPPG with two other MTurkers. In the TPPG, a Dictator and a Recipient each began with 15 cents; the Dictator then choose whether to give her 15 cents to the Recipient, to do nothing, or to take 15 cents from the Recipient; and finally the Sanctioner received 15 cents and chose how much (if any) to spend on punishing the Dictator, with each cent spent by the Sanctioner reducing the Dictator's payoff by 2 cents.

The DG and TPPG were presented using neutral language (e.g. "reduce" the Dictator's earnings, rather than "punish" the Dictator). We predefined the roles participants could take in the DG and the TPPG in order to gather the largest and most informative dataset. In particular, all participants in the DG played in the role of the Dictator, with the Recipient being an MTurker selected at random from a list of worker IDs of participants in prior experiments we ran. In the TPPG, all participants played in the role of Sanctioner, and were matched with a Dictator (recruited as part of another experiment) who chose to take 15 cents from the Recipient (i.e. who acted maximally selfishly).

The experiment was performed in *Qualtrics*, and screenshots of it are included in Appendix B. Once participants accepted the task and entered their worker IDs, a random number generator determined whether the institution questions came before or after the games, and, within the games, whether the DG came before or after the TPPG. Participants were not aware of the existence of subsequent stages, and specific instructions for each stage were only provided at the relevant time. After completing the three stages, participants completed a demographics questionnaire, from which we collected a set of categorical (sex, education, religious affiliation, ethnicity) and continuous (age, social conservatism, fiscal conservatism, and income) variables to use as controls in our analyses.

To assess comprehension, participants were asked to complete a quiz on the rules of the DG and the TPPG, and were told that bonuses would only be paid if answers to the quiz were correct (for the DG, 92.9% of subjects answered all quiz questions correctly; for the TPPG, 65.4%).

2.2 Results & Discussion

We began by examining the relationship between institutional quality and prosociality in the DG (Figure 1). As predicted, we found a significant positive correlation between institutional quality and amount given across all participants, $\beta = 0.081$, $t = 2.17$, $p = 0.03$ (Table 1 col 1), a marginally significant positive correlation when only considering participants who passed the game comprehension questions, $\beta = 0.069$, $t = 1.76$, $p = 0.078$ (Table 1 col 2), and a significant positive correlation when including all participants and controlling for game comprehension, as well as gender, age, education, ethnicity, religion, social conservatism, fiscal conservatism, and

income, β= 0.082, t= 2.03, p= 0.043 (Table 1 col 3). There were no significant interactions between institutional quality and game or survey order (p> 0.54 for both). We note that the positive relationship between institutional quality and DG giving was almost entirely driven by an increase in the likelihood of giving a non-zero amount, rather than an increase in the amount given by those who gave something (see Appendix C).

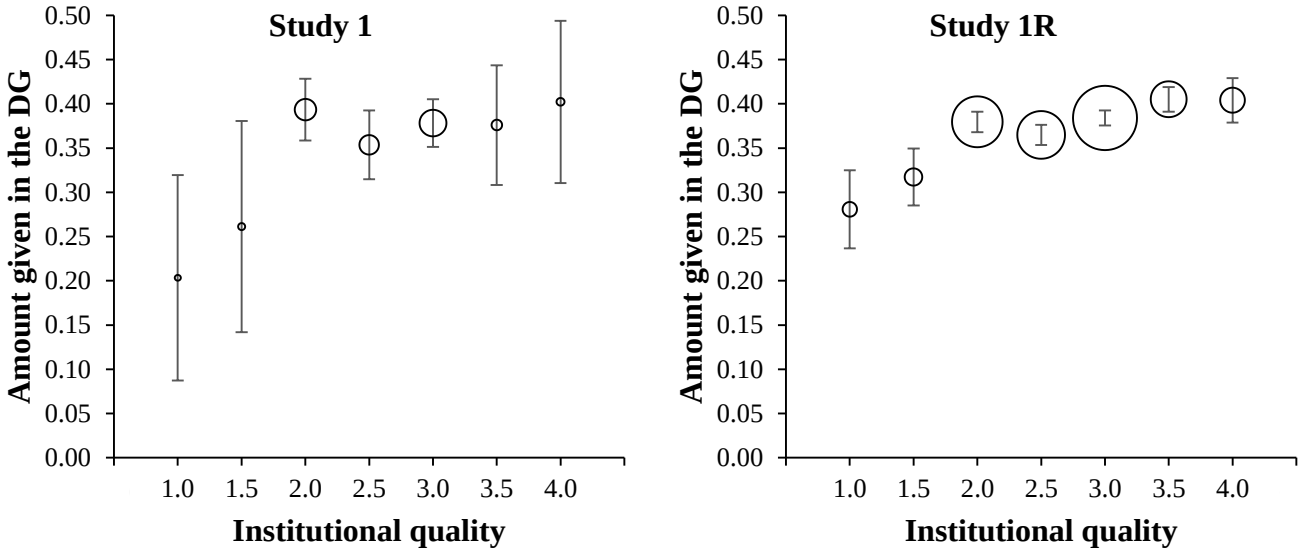


Figure 1. Effect of daily life institutional quality on the amount given in the DG in Study 1 (A) and Study 1R (B) (shown as a fraction of the Dictator's endowment). Institutional quality was binned by rounding up to the nearest 0.5, and average value across all participants in each bin is shown; dot size is proportional to the number of observations in each bin. Error bars indicate 95% confidence intervals.

Because the predicted correlation was only marginally significant when focusing on comprehenders, we ran a replication (Study 1R) to clarify the relationship between institutional quality and dictator giving, with N=1705 (providing 80% power to detect the institutional quality effect observed among comprehenders in Study 1). In Study 1R, participants completed a questionnaire including the questions about confidence in the police and courts from Study 1 and then played a dictator game.[2] Confirming the pattern observed in Study 1, we found in Study 1R that the relationship between institutional quality and amount given in the DG was significant across all participants, β= 0.08, t= 3.29, p= 0.001 (Table 1 col 4), when only considering participants who passed the game comprehension questions, β= 0.058, t= 2.30, p= 0.022 (Table 1 col 5), and when including all subjects and controlling for game comprehension and demographics, β= 0.055, t= 2.08, p= 0.038 (Table 1 col 6). Combining the data from the two studies further

[2] We did not counter-balance order because there were no order effects in Study 1, and we did not include the TPPG or questions about confidence in institutions other than courts and police because the goal of this study was to assess the replicability of the DG relationship reported in Study 1; see Appendix D for details of the experimental design.

supported the significant role of institutional quality in DG giving ($p < 0.002$ for all three comparisons).

Table 1. Linear regression predicting giving in the DG for Study 1 and Study 1R. Shown are raw coefficients (b) and standard errors in parentheses.

	Study 1			*Study 1R*		
	All subjects	*Comprehenders*	*All subjects*	*All subjects*	*Comprehenders*	*All subjects*
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Institutional quality	0.901*	0.764†	0.903*	0.788**	0.574*	0.546*
	(0.416)	(0.433)	(0.444)	(0.239)	(0.25)	(0.263)
Failed comprehension			1.423			0.778
			(1.06)			(0.563)
Female			1.114*			-1.387***
			(0.546)			(0.324)
Age			-0.012			0.026
			(0.027)			(0.014)
Social liberal			0.072			0.123
			(0.225)			(0.133)
Fiscal liberal			-0.251			0.003
			(0.224)			(0.129)
Income			-0.174			-0.023
			(0.122)			(0.068)
Constant	8.700***	8.933***	4.842	9.226***	9.74***	12.861***
	(1.121)	(1.168)	(5.156)	(0.665)	(0.692)	(1.996)
Education dummies	No	No	Yes	No	No	Yes
Ethnicity dummies	No	No	Yes	No	No	Yes
Religion dummies	No	No	Yes	No	No	Yes
N	707	657	707	1705	1555	1705
R^2	0.007	0.005	0.056	0.006	0.003	0.047
Adj R^2	0.005	0.003	0.019	0.006	0.003	0.031

† p<0.1 * p<0.05, ** p<0.01, *** p<0.001.

To provide support for the proposed mechanism underlying this effect, we next performed a mediation analysis using the combined data from Studies 1 and 1R. Our theory argues that high quality institutions hold people accountable and therefore incentivize cooperation, leading people to cooperate in daily life, which in turns leads people to internalize cooperation (and thus cooperate even in settings beyond the reach of the institution). Consistent with this proposal, we found that participants' level of trust in those they interact with in daily life (a proxy for the cooperativeness of the social environment induced by the institutions under which they live) fully mediated the

relationship between institutional quality and dictator giving: using a Sobel-Goodman mediation test with bootstrapping (case resampling, 1000 repetitions), the estimated indirect effect of interpersonal trust on DG giving was significantly different from zero, b= .429, 95% *CI*: .283, .574, while the direct effect of confidence in police and courts on DG giving was not, b= .4, 95% *CI*: -.037, .838. This suggests that higher quality institutions led to higher levels of daily life cooperative interactions, which in turn led to greater internalization of prosociality.

Finally, we examined the relationship between institutional quality and punishment in the TPPG in Study 1 (Figure 2). We did not find a significant correlation between institutional quality and TPP overall, β= 0.011, t=0.28, p= 0.777, when only considering participants who passed the game comprehension questions, β= -0.033, t= -0.71, p= 0.481, or when including all subjects and controlling for game comprehension and demographics, β= 0.001, t=0.02, p= 0.984. We also found no significant interactions between institutional quality and game or survey order (p>0.31 for all).

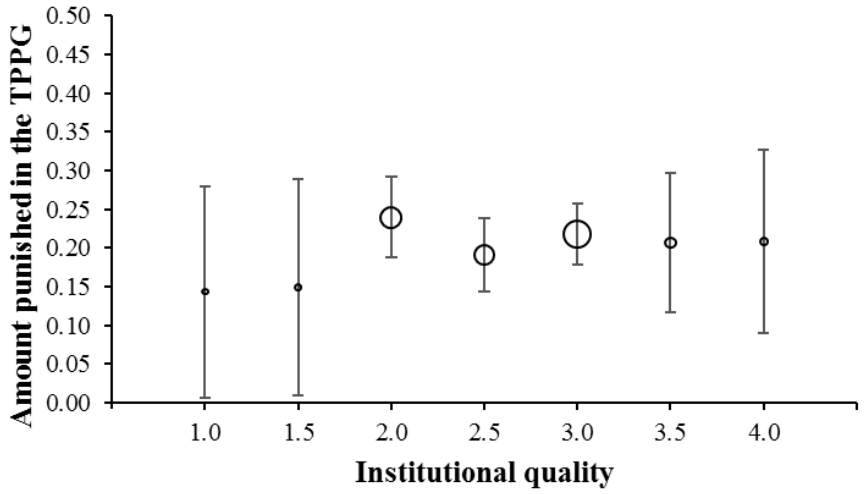


Fig. 2. Effect of daily life institutional quality on the amount punished in the TPPG in Study 1 (shown as fraction of the Sanctioner's endowment). Institutional quality was binned by rounding to the nearest 0.5, and average value across all participants in each bin is shown; dot size is proportional to the number of observations in each bin. Error bars indicate 95% confidence intervals.

Thus, in Study 1 and Study 1R, we observed the predicted correlation whereby stronger institutions were associated with greater prosociality. Conversely, we did not find a significant relationship between institutional quality and norm enforcement in the form of peer punishment. Although these results were consistent with our predictions regarding prosociality, they were only correlational. Study 2 therefore sought to demonstrate causality by experimentally manipulating experienced institutional quality. Study 2 also sought to test the replicability of Study 1's null result regarding the relationship between institutional quality and peer punishment.

3. Study 2: Experimentally manipulating institutional quality

3.1. Materials and methods

3.1.1. Participants

We recruited another 516 MTurkers located in the USA. The average age of participants in this sample was 34 years (min 18, max 68), and 45.9% were female. The task took participants between 10 and 15 minutes to complete and they received a $1 flat fee for participating, plus a variable bonus that on average totaled $1.27 (min $0.00, max $2.80). As in Study 1, we prevented repeat participation by removing duplicate worker IDs and IP addresses.

3.1.2. Method

Study 2 consisted of three stages. The first stage experimentally manipulated institutional quality by having participants play ten rounds of a *public goods game* (PGG) with two other MTurkers in which the effectiveness of a central punishment institution was varied across experimental conditions. In the second and third stages, participants played a single-shot DG and TPPG, respectively, as in Study 1. The PGG stage was always played first, exposing participants to a given level of institutional quality; then, after completing the PGG, participants played the DG and TPPG in random order. Importantly, the DG and TPPG were played with new partners who had not been involved in the prior PGG.

In Study 2, institutional quality was operationalized as the extent to which top-down forces led participants to act in a way that benefited the greater good. This cooperative behavior was measured using a repeated PGG that lasted ten rounds. To ensure that participants did not have varying expectations of the length of the game, the total number of rounds was made public knowledge (as in most repeated PGG experiments). In each round, participants received 140 points and decided how many points (if any) to contribute to a group project All contributions were added up and multiplied by a factor *M*. The resulting number of points was then divided equally between the three participants in the group, irrespective of how much each person contributed. Thus, contributions benefit the group as a whole, but the individual always earned the most by not contributing. Across sessions, we randomized the contribution multiplier *M* to be either 1.2 or 1.5, to test the robustness of our results. Thus, the marginal per capita return (MPCR) that a participant obtained from their contributions was equal to either 0.4 (1.2/3) or 0.5 (1.5/3).

To experimentally manipulate the strength of the enforcement institution, and thus the extent to which participants cooperated in the PGG, we varied the presence and effectiveness of a centralized punishment mechanism. In the control, the PGG was played as described above, with no punishment possible. Across four different treatment conditions, conversely, top-down punishment was added: in each round there was some probability that each of the participants' contributions would be "inspected". If inspected, participants were fined if they had not contributed the maximum amount (they lost 1.5 points for each point below the maximum

contribution of 140 points). The probability of inspection was varied across four levels (5%, 10%, 15%, and 20%). In total, therefore, there were 10 experimental conditions for the Stage 1 repeated PGG: [*M*=1.2, 1.5] x [Control (i.e. no inspection), 5% inspection, 10% inspection, 15% inspection, 20% inspection].

The one-shot DG and TPPG were the same as in Study 1, except that the Dictator had 1400 points to divide in the DG, and each player in the TPPG started with a 1400 point endowment. As in Study 1, we predefined the roles that participants could take in the DG and the TPPG. To avoid issues related to potential income effects, we followed a common practice in experimental economics, informing participants that only one of the three stages would be randomly selected for payment (so that earning more points in one stage should not make participants feel like they had more to spend in subsequent stages).[3] Points earned in the selected stage were converted into dollars at an exchange rate of 10 points equaling $0.01, and specific instructions for each stage were only provided upon completing the prior stage.

As Stage 1 involved live repeated play between multiple participants, the study was conducted using the LIONESS software platform for interactive online experiments (Arechar, Molleman, & Gachter, 2016). Once participants accepted the task on Mturk, they were taken to an introductory page where general instructions emphasized the interactive nature of the task (screenshots of the experiment are also included in Appendix B). To maximize data quality given the longer task and higher per-participant cost, we *required* game comprehension prior to playing each stage: after reading the instructions, participants could not advance to the game until they correctly answered all comprehension questions (they were allowed an unlimited number of attempts). To account for the non-independence of behavior from participants in the same PGG group, all of our regressions clustered standard errors on PGG group. Our regressions predicting contributions in the PGG also clustered on individual, as there were multiple observations per individual.

3.2 Study 2 Results & Discussion

We began with a manipulation check, assessing how PGG contributions varied based on inspection probability (Figure 3). As expected, linear regression revealed a significant positive effect of inspection probability on contributions, β=0.296, t=6.36, p<0.001 (Table 2 col 1). As can be seen in Figure 3, however, this relationship was strongly non-linear (Table 2 col 2). Contributions were lower in the control compared to the punishment conditions (i.e. collapsing across 5%, 10%, 15%, and 20% inspection probabilities), β=0.371, t=7.39, p<0.001 (Table 2 col 3); but contributions did not vary across the different punishment conditions (p>0.11 for linear and quadratic models; Table 2 col 4 and 5).

[3] See Cox, Sadiraj, and Schmidt (2014) for a discussion on the issue and Lee (2008) for an application of the method.

Table 2. Linear regression with robust standard errors clustered on group, predicting PGG contribution across all rounds in Study 2. Shown are raw coefficients (b) and standard errors in parentheses.

	All conditions			*Punishment conditions (inspection>0)*	
	(1)	*(2)*	*(3)*	*(4)*	*(5)*
Inspection probability	1.688***	4.742***		0.309	-2.597
	(0.265)	(0.949)		(0.300)	(1.892)
Inspection probability^2		-15.439***			11.544
		(4.213)			(7.287)
Inspection>0 dummy			0.367***		
("Institutional quality")			(0.050)		
Constant	0.581***	0.508***	0.460***	0.788***	0.936***
	(0.037)	(0.044)	(0.046)	(0.044)	(0.103)
N	5160	5160	5160	4050	4050
R^2	0.087	0.113	0.138	0.002	0.009

* p<0.05, ** p<0.01, *** p<0.001.

Thus only a minimal incentive was needed to secure consistent PGG cooperation in our experiment, with 5% inspection stabilizing to the same extent as 20% inspection. This is particularly interesting given that from a strictly economic perspective, even a 20% inspection probability with a 1.5x fine was too small to make full cooperation the payoff-maximizing choice. We also saw that these effects were robust to MPCR: although there was a significant positive main effect of MPCR, β= 0.096, t= 2.06, p= 0.04, which is not surprising due to the increased incentive to cooperate, there was no significant interaction between MPCR level (0.4 versus 0.5) and inspection probability (p>0.50 for both linear and quadratic models).

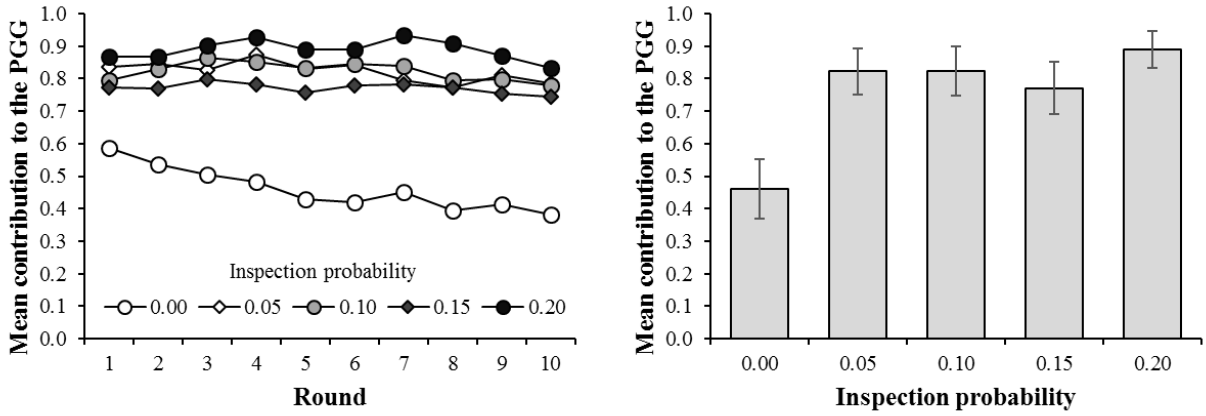


Fig. 3. Contributions in the public goods game, based on inspection probability. (a) By PGG round. (b) Averaged over all 10 rounds; error bars indicate 95% confidence intervals.

Thus adding centralized punishment (i.e. inspection probability greater than zero) increased contributions relative to the control (nearly doubling contribution rates), but did so to

such an extent that further increases in inspection probability had little effect on contributions (perhaps due to a ceiling effect). Furthermore, the presence of centralized punishment increased cooperation to same extent for both MPCRs. Therefore, because changing the institution was only predicted to affect prosociality and punishment in so much as it altered PGG contributions, our analyses of subsequent DG and TPPG play compared behavior in the control ("low institutional quality") to behavior in the presence of centralized punishment (i.e. collapsing across the four non-zero inspection probability conditions; "high institutional quality"), and collapsed across both MPCRs.

What, then, was the effect of experimentally induced institutional quality on subsequent prosociality and punishment? We begin with giving in the DG (Figure 4). Consistent with our predictions and the correlational results from Study 1, we observed a significant positive relationship between institutional quality and average DG giving, β= 0.111, t= 2.48, p= 0.014 (Table 3 col 1; low quality institution: 25.3% of endowment given; high quality institution, 32.4% of endowment given).[4] Also, as in Study 1, we observe a significant positive relationship between institutional quality and the decision to give in the DG, but not with the amount given conditional on giving (for full analysis see Appendix C).

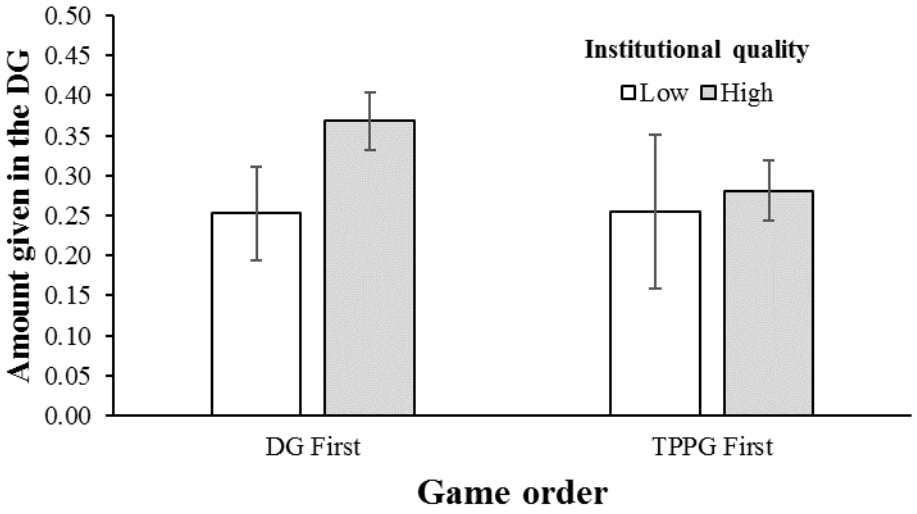


Fig. 4. Effect of experimentally induced institutional quality on the amount given in the DG in Study 2. Error bars indicate 95% confidence intervals.

[4] In support of our decision to collapse across non-zero inspection probabilities, there was a significant non-linear relationship between average DG giving and un-collapsed inspection probability, linear term: β= 0.408, t= 2.59, p= 0.010; quadratic term: β= -0.379, t= -2.38, p= 0.018; with no significant relationship between DG giving and inspection probability when restricting to non-zero inspection probabilities (linear model: β= -0.06, t= -1.15, p= 0.251; non-linear model: linear term, β=0.076, t= 0.26, p= 0.792, quadratic term, β= -0.138, t= -0.47, p= 0.636).

Furthermore, although the interaction between institutional quality and game order (i.e. whether the DG or TPPG was played first) did not reach statistical significance, β= -0.167, t= -1.54, p= 0.126 (Table 3 col 2), there was some evidence of an interaction: in the "cleaner" order where the DG came first (and thus was not influenced by the TPPG), we see a large and significant treatment effect, β= 0.195, t= 3.50, p= 0.001 (low quality institution: 25.2% of endowment given; high quality institution: 36.8% of endowment given). Conversely, there was no significant treatment effect in the somewhat confounded order where the TPPG came before the DG, β= 0.039, t= 0.55, p= 0.585 (low quality institution: 25.5% of endowment given; high quality institution: 28.1% of endowment given).

Table 3. Linear regression with robust standard errors clustered on group, predicting dictator game giving and third-party punishment in Study 2. Shown are raw coefficients (b) and standard errors in parentheses.

	DG giving		Third-party punishment	
	(1)	*(2)*	*(3)*	*(4)*
Institutional quality	0.071*	0.115**	0.038	0.112*
	(0.028)	(0.033)	(0.038)	(0.054)
Order		0.002		0.090
		(0.052)		(0.065)
Institutional quality x Order		-0.089		-0.155*
		(0.058)		(0.075)
Constant	0.254***	0.252***	0.246***	0.204***
	(0.025)	(0.028)	(0.033)	(0.046)
N	516	516	516	516
R^2	0.012	0.034	0.002	0.011

* p<0.05, ** p<0.01, *** p<0.001.

Finally, we turn to norm enforcement in the TPPG (Figure 5). As in Study 1, we found no significant relationship between institutional quality and average punishment, b= 0.042, t=0.99, p= 0.324 (Table 3 col 3; low quality institution: 25.6% of endowment spent on punishment; high quality institution, 28.4% of endowment spent on punishment).[5] We did, however, find a significant interaction between institutional quality and game order, β= -0.202, t= -2.05, p= 0.041 (Table 3 col 4), such that in the cleaner order where the TPPG came before the DG, there was no significant treatment effect, β= -0.048, t= -0.81, p= 0.418, while in the somewhat confounded order where the DG came before the TPPG, there was a significant positive relationship between institutional quality and TPP, β= 0.122, t= 2.07, p= 0.042.

[5] This lack of significant relationship was not an artifact of collapsing across non-zero inspection probabilities: there was neither a linear, β=0.049, t= 1.09, p= 0.278, nor non-linear (linear term: β= 0.096, t= 0.64, p= 0.525, quadratic term: β= -0.05, t= -0.32, p= 0.750) effect of uncollapsed inspection probability on average punishment.

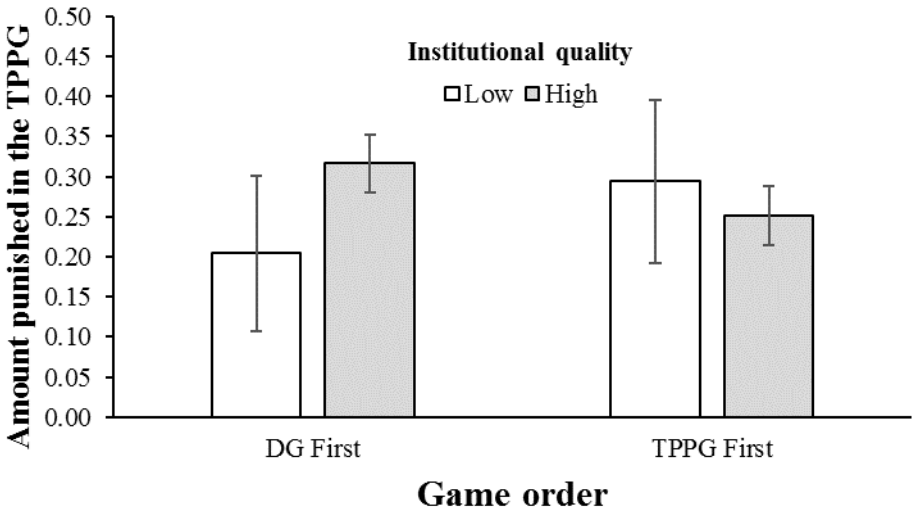


Fig. 5. Effect of experimentally induced institutional quality on amount spent on punishment in the TPPG in Study 2. Error bars indicate 95% confidence intervals.

This order effect, where treatment only influenced punishment in the "DG First" order where the DG immediately preceded the TPPG, suggests that whatever influence the institutional quality manipulation had on punishment was *indirect*, via changes in DG giving. Accordingly, DG giving fully mediated the relationship between institutional quality condition and punishment in the TPPG, as shown by a mediational model using the bootstrapping method (Hayes, 2013). In this model, institutional quality condition was the independent variable (X); DG giving was the mediator (M); and TPPG punishment was the dependent variable (Y). The mediational model was estimated with 1,000 iterations sampled at the level of the PGG group using *sgmediation* in STATA. The estimated indirect effect of DG giving on TPPG punishment was 0.030 (SE = .013) with a 95% confidence interval of 0.005 to 0.055. The confidence interval did not include zero, indicating that DG giving significantly mediated the effect of institutional quality condition on TPPG punishment. People in the no-punishment control gave less in the DG, and in turn, punished less in the TPPG. The estimated direct effect of institutional quality condition on TPPG punishment was 0.008 (SE = .038) with a 95% confidence interval of -0.066 to 0.082. The confidence interval of the direct effect included zero, indicating that DG giving fully mediated the relationship between institutional punishment condition and TPPG punishment.

Taken together, the results of Study 2 therefore indicated that exposure to a low quality institution that allowed selfishness and thus incentivized free-riding in the PGG (i) led directly to people engaging in less "pure" prosociality which will not bring them future benefits, but (ii) only made people less inclined to punish selfishness *indirectly* by decreasing prosociality (and thus leading to less punishment of others' selfishness).

General Discussion

Here we have presented evidence from two studies exploring the impact that top-down institutional incentives to cooperate can have on prosociality and punishment. Not only do such incentives increase cooperation directly, but they also influence prosociality in situations beyond the reach of the institution. We have also presented some insight into boundary conditions for this effect: we did not find evidence that institutional quality directly affected norm enforcement, but instead found some evidence of an indirect effect on punishment via increasing prosociality.

With respect to prosociality, in Study 1, we presented correlational evidence (based on self-reported confidence in the police and courts) that living under stronger institutions was associated with somewhat greater prosociality. These findings regarding prosociality were consistent with prior work examining cross-cultural correlations between other forms of institutional quality (religion and market integration) and prosociality (Henrich et al., 2010), and extend these prior findings in several ways. First, we used a measure of institutional quality that directly captured the effectiveness of institutions at materially incentivizing cooperation (and disincentivizing exploitation) – the courts and the police. We also showed that it was possible with this measure to find substantial variation, even among residents of the United States participating online – although it seems likely that we would have observed a stronger relationship had we used a more diverse (e.g. less Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010)) sample in which there was more underlying variation in institutional exposure. Future work should therefore attempt to replicate these spill-over results using cross-cultural samples.

In Study 2, we assessed the robustness of these results and addressed the question of causality by experimentally manipulating institutional quality (via a centralized punishment institution) in a repeated PGG. We found that an institution that held individuals accountable for selfishness positively affected prosociality in a novel context (i.e. incentives to cooperate in the PGG increased subsequent giving in the DG), replicating the correlational finding from Study 1.

Thus we have provided convergent evidence that the quality of the institutions one is exposed to "spills over" to affect prosociality in other contexts. We find indications that this generalization can be the product of long term exposure (daily-life institutions in Study 1), or can occur over a short timeframe (experimentally induced institutions in Study 2). Exploring the connection between these different timescales, and whether they operate via the same cognitive mechanisms, is a promising direction for future study. Relatedly, because Study 2 did not include a baseline condition (or a pre-treatment measure of prosociality), we cannot tell the extent to which the high quality institution increased prosociality versus the low quality institution decreasing prosociality. This issue should also be addressed in future work, as should quantifying how different lengths of exposure translate into different effect sizes. More generally, it is important for future work to explore the generalizability of our results using more real-world measures, such as natural experiments (exploiting variation in institutional quality across locations) or actual field experiments involving randomization (e.g. Kraft-Todd et al., 2015).

Our Study 2 results regarding induced PGG cooperation spilling over to prosociality in the DG are consistent with the results of Peysakhovich and Rand (2016), who found that playing long versus short repeated PD games (leading to high versus low levels of bilateral cooperation) spilled over to DG giving (as well as other 1-shot cooperation games). Thus we show that their results using repeated interactions between pairs of people extend to formal top-down institutional punishment, and to group cooperation between more than two people. Furthermore, in their cooperation stage, participants had to learn over time to cooperate or defect – in the first decision of the repeated PD stage, cooperation rates were very similar between the long and short repeated game conditions. In our Study 2, conversely, PGG contribution rates in the very first round were already much higher in the high institutional quality conditions compared to the low institutional quality conditions (although the difference between conditions did increase over time). This immediate difference, together with the subsequent prosocial spillovers, suggests that changes in prosociality via habituation can occur even when the initial behavioral change is the result of deliberate processes. Future work should investigate in more detail how this habituation occurs.

Our results regarding prosocial spillovers in Study 2 are also consistent with the findings of Falkinger, Fehr, Gächter, and Winter-Ebmer (2000), who observed that the strong positive effect of a sanctioning institution on PGG contributions spilled over into the initial rounds of a second stage where participants played the same PGG but with the sanctioning institution removed (although the spillover effect they observed was only transient, and contributions decreased relatively quickly in the absence of institutional incentive). In having their participants play the same game in both stages, however, it was unclear whether their observation was truly the result of a change in participants' prosociality, or merely the result of a more mundane rote learning process (e.g. just pressing the same button repeatedly, or continuing to choose from the same end of the range of possible options). Our experiment did not have this issue, as we measured prosocial spillovers between different games, from a PGG (where players had a 140 unit endowment and most players picked the maximum contribution amount of 140) to a DG (where players had a 1400 unit endowment and most players gave half, 700 units). Thus we demonstrate true spillover effects into novel situations, which cannot be explained by simple rote learning. The question of how long these spillovers last, however, remains an important direction for future research.

Interestingly, we found a dissociation between people's decisions about *whether* to give, and *how much* to give conditional on giving: institutional quality was positively related to the former, but unrelated to the latter. This pattern suggests that people hold some notion of what the "right" way to act is (e.g. giving half in the DG), and the institution shapes people's willingness to act on this knowledge. This is in contrast to institutions shifting people's understandings of what is right or wrong in a gradual, graded fashion (e.g. increasingly strong institutions motivating people to give 0% in the DG, then 10%, then 20%, etc.); and is consistent with the observation from a large meta-analysis of DG giving that a majority of people give either nothing or half (Engel, 2011), as well as developmental evidence that among high SES American children, the

probability of giving increases between 3 and 6 years of age while the amount given among givers does not change (Blake & Rand, 2010).

The fact that we do observe spillovers from multilateral cooperation in the PGG to unilateral donation in the DG raises the question of how our results relate to recent findings regarding promoting intuition versus deliberation in the DG. On the one hand, our findings are consistent with the spillover predictions of the SHH, given the observation that PGG and DG play correlate within an individual and both reflect an underlying "cooperative phenotype" (Peysakhovich et al., 2014) – such that increasing prosociality in one domain should spill over to prosociality in other domains. On the other hand, it has been argued that the SHH predicts that only women should intuitively give in unilateral altruism settings like the DG (Rand, Brescoll, Everett, Capraro, & Barcelo, 2016). This is because in daily life, women are expected to be altruistic (and punished for not doing so) to a much greater extent than men (Eagly, 1987; Heilman & Okimoto, 2007), such that altruism is only typically long-run payoff maximizing for women. Consistent with this argument, a 22 study meta-analysis of the effect of promoting intuition versus deliberation on DG giving found that intuition only favored DG giving among women and not among men (Rand et al., 2016).[6] We, however, find no interaction between institutional quality and gender in either of our studies (see Appendix E). Thus we are left with the question of why incentivizing cooperation in our lab environment (as well as that of Peysakhovich and Rand (2016)) increases DG giving for both men and women, whereas Rand et al. (2016) find that promoting intuition (which the SHH argues implements the behavior that is typically incentivized in daily life) only increases female DG giving. One possible explanation is that the intuition manipulations applied to the DG in Rand et al. (2016) pushed subjects to see the DG as an instance of daily-life altruism in particular (resulting in women but not men increasing giving); whereas in our studies (where there was no intuition manipulation), participants tended to instead see the DG as an abstract economic game involving prosociality (rather than altruism in particular) - which then facilitated spillover from general moral principles instilled by institutional enforcement (Study 1) and between different games that both involving paying costs to benefit others (Study 2). Testing this idea, and more generally understanding this important difference between studies that promote intuition and studies that examine spillovers, is a key direction for future theoretical and empirical work.

Our findings, as well as other work using peer-based reputational incentives (Nakashima et al., 2016; Peysakhovich & Rand, 2016), run counter to the large body of evidence regarding "crowding out" effects, whereby internal drives to achieve some goal can be supplanted by external incentives which are unrelated to the initial relationship between goal and natural (intrinsic) reward (Deci et al., 1999). Such effects have been specifically demonstrated in the context of prosociality (Frey & Jegen, 2001), raising the question of why we did not observe evidence of top-down institutional punishment of selfishness crowding out subsequent DG prosociality in Study 2. One

[6] This finding stands in contrast to multilateral cooperation, where meta-analysis indicates that intuition favors cooperation for both women and men (Rand, 2016a).

possible explanation is that our intrinsically motivated stage (DG) was sufficiently different from the stage in which the extrinsic motivation was applied (PGG) to avoid crowding out (although the findings of Falkinger et al. (2000) described above suggest that spillovers may still occur even when both stages involve the same behavior). Another possibility is that our institutional mechanism was sufficiently "gentle" to avoid supplanting participants' intrinsic desire to give (while still being strong enough to motivate PGG contributions). Future work should investigate this latter possibility by testing whether a harsher institutional mechanism (i.e. one with a higher inspection probability, or a greater fine for non-contribution) undermines the spillovers we observed and instead leads to crowding out.

Turning from prosociality to punishment, we find much less evidence of spillover effects. In Study 1, we did not find a significant correlation between TPP and self-reported confidence in the police and the courts. This null result stands in contrast to some prior cross-cultural findings (e.g. Henrich et al. (2010)), but is consistent with other work that found cross-cultural variation only in anti-social punishment of cooperators, rather than punishment of selfishness (Herrmann et al., 2008). It is possible, of course, that if we had examined participants from a wider range of backgrounds (e.g. from numerous different countries, rather than just the United States), we might have observed a similar correlational result regarding punishment to that of Henrich et al. (2010).

Similarly, Study 2 did not find an overall effect of an accountability-inducing institution on participants' TPP. This lack of positive overall spillover is broadly consistent with the even more extreme findings of Romaniuc, et al. (2016), who observed that institutional punishment actually led to a decrease in subsequent peer punishment in a repeated PGG. Our Study 2 did, however, find an order effect in which the institution treatment increased TPP when participants play a DG immediately between the TPPGs. This order effect, together with the observation that the treatment effect on TPP in this order was eliminated when controlling for DG giving, suggests that the institution treatment had an *indirect* effect on punishment via increasing prosociality. This order effect also reconciles our results with those of Peysakhovich and Rand (2016), who *did* observe spillovers from repeated PD play to third-party punishment. In their design, however, participants made decisions in both roles of the TPPG, and, critically, always made their Dictator decision immediately before making their Sanctioner decision – similarly to our DG First order (in which we also observed a treatment effect).

The fact that we did not observe a direct effect of institutional quality on punishment sheds important light on the mechanism underlying the effect on DG giving that we did observe. In particular, the lack of direct effect on punishment suggests that institutional quality was not impacting prosociality via a change in perceived social norms. If exposure to high versus low quality institutions impacted prosociality by changing people's explicit understanding of what behavior is appropriate (i.e. their perception of the social norm), this would have also led to changes in punishment in the TPPG: punishment is (at least in part) driven by anger arising from norm violations (Fehr & Fischbacher, 2004; Jordan et al., 2015), such that changes in one's understanding of the social norm would lead to changes in punishment. Instead, the lack of direct

punishment effect is consistent with the SHH's focus on the spillover of advantageous behavior: by this logic, institutions that incentivize prosociality should lead to more subsequent prosociality, but should not influence subsequent punishment (as such punishment was not incentivized by the institution).

Finally, our Study 2 punishment results suggest that correlational findings linking institutional quality and third-party punishment of selfishness (e.g. Henrich et al. (2010)) may not reflect an actual direct relationship, but instead an indirect effect. This indirect effect could, for example, be driven by hypocrisy concerns: individuals who act selfishly in the DG may avoid subsequently punishing selfishness in the TPPG because doing so would be hypocritical. The fact that punishment spillovers only occurred in the DG First order is not consistent, conversely, with an indirect effect driven by people trying to signal their trustworthiness by punishing – signaling-based punishment has been shown to be *reduced* by the concurrent opportunity to signal trustworthiness via DG giving (Jordan et al., 2016). Further investigation of the basis of this indirect spillover effect is an important direction for future work.

In sum, we have provided evidence of the role that formal institutions which punish bad behavior play in shaping human prosociality. These findings shed light on why we are often willing to incur costs to benefit strangers (even when there will be no future consequences for our behavior), and also help to explain why the extent of such prosocial behavior varies markedly cross-culturally. The institutions that govern our lives help to shape our willingness to choose "right" over "wrong."

Acknowledgements

Financial support from the John Templeton Foundation is gratefully acknowledged, as is helpful discussion and comments from Adam Bear, Alex Peysakhovich, and three anonymous reviewers. The authors declare that all relevant data and measures used in this study have been included in this manuscript.

References

Anderson, E. (2000). *Code of the street : decency, violence, and the moral life of the inner city* (1st pbk. ed.). New York: W.W Norton.

Arechar, A. A., Molleman, L., & Gachter, S. (2016). Conducting interactive experiments online. *Mimeo*.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(4489), 1390-1396.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325-344.

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology, 40*(2), 290.

Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences, 113*(4), 936-941.

Blake, P. R., McAuliffe, K., Corbit, J., Callaghan, T. C., Barry, O., Bowie, A., . . . Warneken, F. (2015). The ontogeny of fairness in seven societies. *Nature, 528*(7581), 258-261.

Blake, P. R., & Rand, D. G. (2010). Currency value moderates equity preference among young children. *Evolution and Human Behavior, 31*(3), 210-218.

Campbell, J. K. (1966). Honor and the Devil. In J. G. Péristiany (Ed.), *Honour and shame: the values of Mediterranean society* (pp. 265 p.). Chicago: University of Chicago Press.

Campbell, R., & Sowden, L. (1985). *Paradoxes of rationality and cooperation : prisoner's dilemma and Newcomb's problem*. Vancouver: University of British Columbia Press.

Cappelen, A. W., Moene, K. O., Sørensen, E. Ø., & Tungodden, B. (2013). Needs Versus Entitlements - An International Fairness Experiment. *Journal of the European Economic Association, 11*(3), 574-598. doi:10.1111/jeea.12000

Capraro, V., & Cococcioni, G. (2015). Social setting, intuition, and experience in laboratory experiments interact to shape cooperative decision-making. *Proc Roy Soc B*, doi:10.1098/rspb.2015.0237.

Capraro, V., Jordan, J. J., & Rand, D. G. (2014). Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments. *Scientific Reports, 4*, 6790.

Chakroff, A., & Young, L. (2014). The prosocial brain: perceiving others in need, and acting on it. In L. M. Padilla-Walker & G. Carlo (Eds.), *Prosocial Development : A Multidimensional Approach* (pp. 90-111): Oxford Scholarship Online.

Chudek, M., & Henrich, J. (2011). Culture gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in cognitive sciences, 15*(5), 218-226.

Cohen, D., Bowdle, B. F., Nisbett, R. E., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An "Experimental ethnography". *Journal of Personality and Social Psychology, 70*(5), 945-960.

Cox, J. C., Sadiraj, V., & Schmidt, U. (2014). Paradoxes and mechanisms for choice under risk. *Experimental Economics, 18*(2), 215-250. doi:10.1007/s10683-014-9398-8

Cushman, F., & Macindoe, O. (2009). *The coevolution of punishment and prosociality among learning agents.* Paper presented at the Proceedings of the 31st annual conference of the cognitive science society.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin, 125*(6), 627-668. doi:10.1037/0033-2909.125.6.627

Dreber, A., Fudenberg, D., & Rand, D. G. (2014). Who Cooperates in Repeated Games: The Role of Altruism, Inequity Aversion, and Demographics. *Journal of Economic Behavior & Organization, 98*, 41-55.

Eagly, A. H. (1987). *Sex Differences in Social Behavior: A Social-role Interpretation*. Mahwah, New Jersey: L. Erlbaum Associates.

Edgerton, R. B. (1971). *The individual in cultural adaptation; a study of four East African peoples*. Berkeley,: University of California Press.

Engel, C. (2011). Dictator Games: A Meta Study. *Experimental Economics, 14*, 583-610.

Espín, A. M., Brañas-Garza, P., Herrmann, B., & Gamella, J. F. (2012). Patient and impatient punishers of free-riders. *Proceedings of the Royal Society B: Biological Sciences, 279*(1749), 4923-4928. doi:10.1098/rspb.2012.2043

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol., 59*, 255-278.

Falkinger, J., Fehr, E., Gächter, S., & Winter-Ebmer, R. (2000). A simple mechanism for the efficient provision of public goods: Experimental evidence. *American Economic Review*, 247-264.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63-87.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in Simple Bargaining Games. *Games and Economic Behavior, 6*, 347-369.

Frey, B. S., & Jegen, R. (2001). Motivation Crowding Theory. *Journal of Economic Surveys, 15*(5), 589-611. doi:10.1111/1467-6419.00150

Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1553), 2651-2661. doi:10.1098/rstb.2010.0135

Galinsky, A., & Schweitzer, M. (2015). Friend & foe: When to cooperate, when to compete, and how to succeed at both: New York: Crown Business.

Gneezy, U., & Rustichini, A. (2000). Pay Enough or Don't Pay At All. *Quarterly Journal of Economics, 115*(3), 791-810.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557-559.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.*: Guilford Press.

Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks?: The implied communality deficit. *Journal of Applied Psychology, 92*(1), 81-92. doi:10.1037/0021-9010.92.1.81

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., . . . Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and brain science, 28*, 795-855.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2010). Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. *Science, 327*(5972), 1480-1484. doi:10.1126/science.1182238

Herrmann, B., Thoni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362-1367.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The Online Laboratory: Conducting Experiments in a Real Labor Market. *Experimental Economics, 14*(3), 399-425. doi:10.1007/s10683-011-9273-9

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*.

Jordan, J. J., McAuliffe, K., & Rand, D. G. (2015). The Effects of Endowment Size and Strategy Method on Third-Party Punishment. *Experimental Economics*, doi:10.1007/s10683-10015-19466-10688.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697-720.

Kiyonari, T., Tanida, S., & Yamagishi, T. (2000). Social exchange and reciprocity: confusion or a heuristic? *Evolution and Human Behavior, 21*, 411-427.

Kraft-Todd, G., Yoeli, E., Bhanot, S., & Rand, D. (2015). Promoting cooperation in the field. *Current Opinion in Behavioral Sciences, 3*, 96-101. doi:http://dx.doi.org/10.1016/j.cobeha.2015.02.006

Lee, J. (2008). The effect of the background risk in a simple chance improving decision model. *Journal of Risk and Uncertainty, 36*(1), 19-41. doi:10.1007/s11166-007-9028-3

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis. *Journal of Personality and Social Psychology, 28*(1), 129.

Nakashima, N. A., Halali, E., & Halevy, N. (2016). Third parties promote cooperative norms in repeated interactions. *Journal of experimental social psychology*. doi:10.1016/j.jesp.2016.06.007

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1514483113

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans Display a 'Cooperative Phenotype' that is Domain General and Temporally Stable. *Nature Communications, 5*, 4939. doi:10.1038/ncomms5939

Peysakhovich, A., & Rand, D. G. (2016). Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory. *Management Science, 62*(3), 631–647.

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution, 30*(2), 98-103. doi:http://dx.doi.org/10.1016/j.tree.2014.12.003

Raihani, N. J., Mace, R., & Lamba, S. (2013). The Effect of $1, $5 and $10 Stakes in an Online Dictator Game. *PLoS ONE, 8*(8), e73131. doi:10.1371/journal.pone.0073131

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology & Evolution, 27*(5), 288-295.

Rand, D. G. (2016a). Cooperation (Unlike Altruism) is Intuitive for Men as Well as Women. *Available at SSRN: http://ssrn.com/abstract=2722981*.

Rand, D. G. (2016b). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*. doi:10.1177/0956797616654455

Rand, D. G., Brescoll, V. L., Everett, J. A. C., Capraro, V., & Barcelo, H. (2016). Social heuristics and social roles: Intuition favors altruism for women but not for men. *Journal of Experimental Psychology: General, 145*(4), 389-396.

Rand, D. G., & Nowak, M. A. (2013). Human Cooperation. *Trends in cognitive sciences, 17*(8), 413-425.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Green, J. D. (2014). Social Heuristics Shape Intuitive Cooperation. *Nature Communications, 5*, 3677.

Sapienza, P., Zingales, L., & Guiso, L. (2006). *Does culture affect economic outcomes?* Retrieved from

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin, 119*(1), 3.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences, 23*(05), 701-717.

Titmuss, R. (1970). The gift relationships. *London, George Allen*.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation. *Current Anthropology, 53*(6), 673-692.

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and sociobiology, 11*(4), 375-424.

Van Lange, P. A. M., De Bruin, E., Otten, W., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. *Journal of Personality and Social Psychology, 73*(4), 733.

World Values Survey. (2012). Retrieved November 2012, from World Values Survey Association http://www.worldvaluessurvey.org

Yamagishi, T., Hashimoto, H., & Schug, J. (2008). Preferences Versus Strategies as Explanations for Culture-Specific Behavior. *Psychological Science, 19*(6), 579-584. doi:10.1111/j.1467-9280.2008.02126.x

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., . . . Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1212126109

Appendix A

Here we present a robustness check on the Study 1 analyses looking at the relationship between cooperation and punishment and the quality of the institutions in participants' daily lives. Specifically, we use the full six-item institution confidence scale from the World Values Survey, which asks about police, courts, government, political parties, civil services, and banking industry (whereas the main text uses only the police and courts items).

As in the main text Section 2.2, we found a positive correlation between institutional quality and giving overall, β= 0.081, t= 2.15, p= 0.032 (Figure A1), among comprehenders only, β= 0.064, p= 0.099, and including all participants and controlling for game comprehension as well as demographics, β= 0.077, t= 1.96, p= 0.051. No significant interactions were found between institutional quality and the order in which the stages were presented (p>0.16 for all).

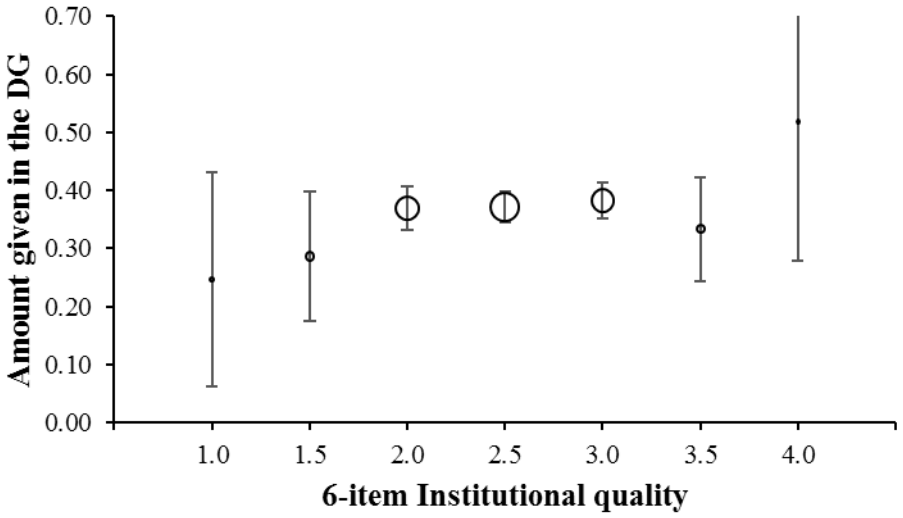


Fig. A1. Effect of daily life institutional quality on amount given in the DG (shown as fraction of the Dictator's endowment) in Study 1. Institutional quality was binned by rounding to the nearest 0.5, and average value across all participants in each bin is shown; dot size is proportional to the number of observations in each bin. Error bars indicate 95% confidence intervals.

Next we examined punishment in the TPPG using the 6-item institutional quality scale. As in the main text, we did not find a significant relationship between institutional quality and average punishment, either without controls, β= 0.045, t= 1.19, p= 0.234 (Figure A2), or including them, β= 0.041, t= 1.05, p= 0.30; or when excluding those participants who failed game comprehension, β= -0.001, t= -0.03, p= 0.979. We also found no significant interactions between institutional quality and either game or survey order (p>0.28 for all).

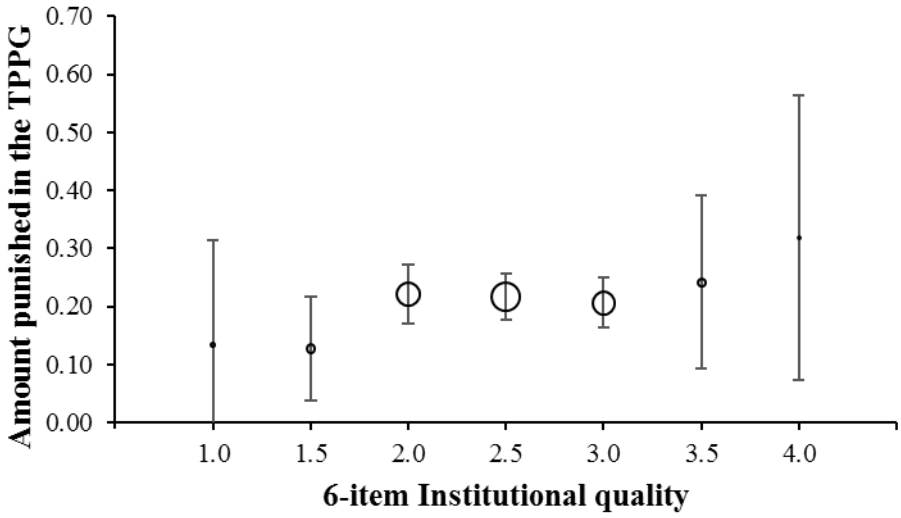
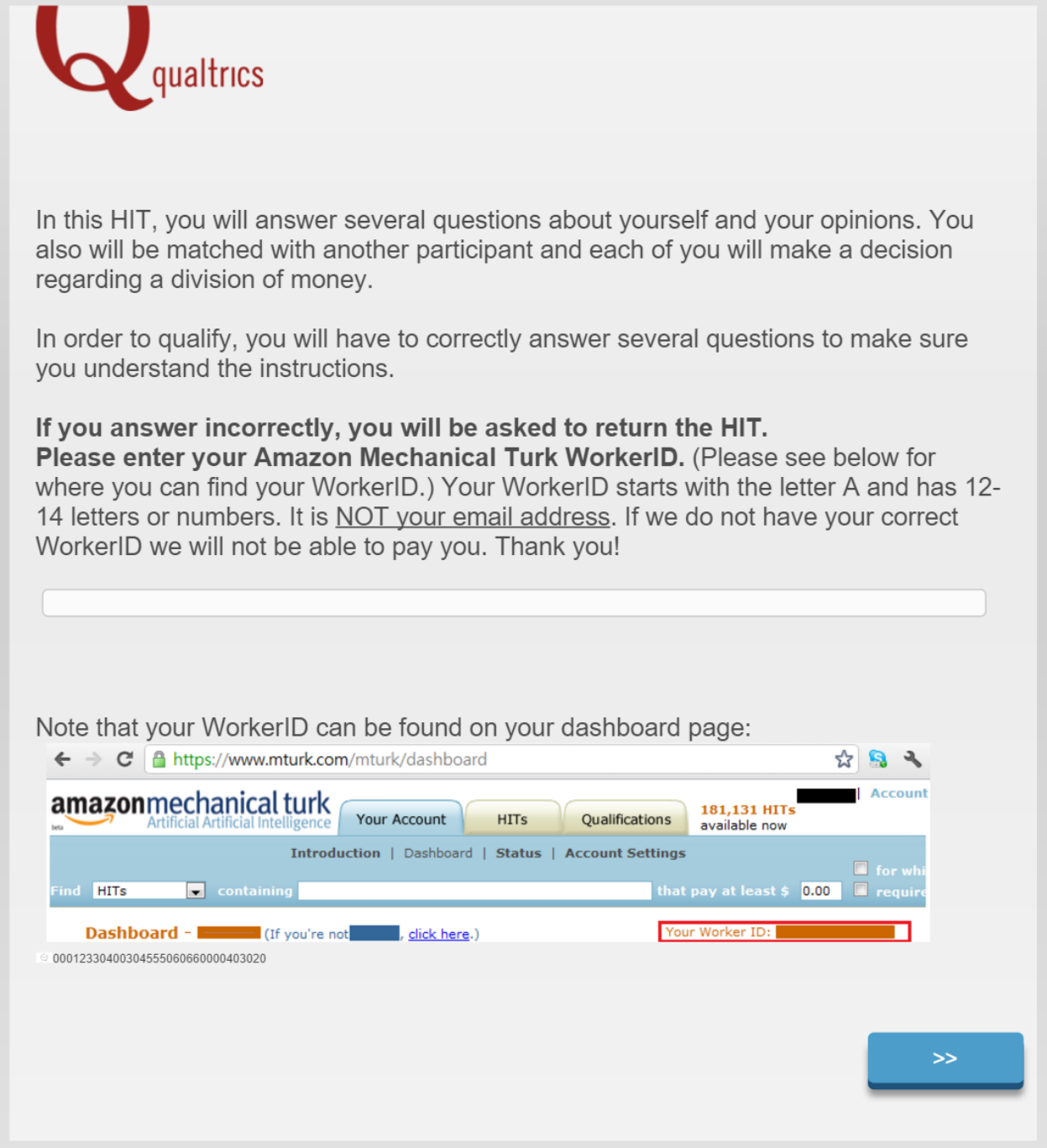


Fig. A2. Effect of daily life institutional quality on amount spent on punishing in the TPPG (shown as fraction of the Sanctioner's endowment) in Study 1. Institutional quality was binned by rounding to the nearest 0.5, and average value across all participants in each bin is shown; dot size is proportional to the number of observations in each bin. Error bars indicate 95% confidence intervals.

Appendix B

Here we present screenshots of Studies 1 and 2, as seen by the participants.

Study 1




qualtrics
Click to write the question text
MOFFAT CRES
Type the text
reCAPTCHA
Privacy & Terms
>>

Please read the directions carefully.

In this part of the HIT you will play a two person game where you have been randomply assigned to interct with one other Mturk worker. You have been assigned to be in Role A and another person from MTurk will be in Role B.

Both of you will receive the same set of intstrucions. Once you make your decision you will have no other interaction with this player.
You can only play this game once.

In addition to the payment you each receive for participating in this HIT, you can earn more as a **bonus**, as follows:

You will get to make an offer of how to split a bonus of 30 cents between you and the person in Role B. The person in Role B must accept this offer, so you and the other person will be paid a bonus according to your proposed split. Once you make your decision, the game is over and you and the Role B person will have no other opportunities to affect each other's bonuses.

Please answer these practice questions below.

If you offer 5 out of 30 cents to the Role B person, what bonus will the Role B person get (in cents)?

- 15 cents
- 5 cents
- 10 cents

Both of you will receive the same set of intstrucions. Once you make your decision you will have no other interaction with this player.
You can only play this game once.

In addition to the payment you each receive for participating in this HIT, you can earn more as a **bonus**, as follows:

You will get to make an offer of how to split a bonus of 30 cents between you and the person in Role B. The person in Role B must accept this offer, so you and the other person will be paid a bonus according to your proposed split. Once you make your decision, the game is over and you and the Role B person will have no other opportunities to affect each other's bonuses.

Please answer these practice questions below.

If you offer 5 out of 30 cents to the Role B person, what bonus will the Role B person get (in cents)?

- 15 cents
- 5 cents
- 10 cents

If you offer 20 out of 30 cents to the Role B person, what bonus will you get (in cents)?

- 30 cents
- 20 cents
- 10 cents

>>

Both of you will receive the same set of intstrucions. Once you make your decision you will have no other interaction with this player.
You can only play this game once.

In addition to the payment you each receive for participating in this HIT, you can earn more as a **bonus**, as follows:

You will get to make an offer of how to split a bonus of 30 cents between you and the person in Role B. The person in Role B must accept this offer, so you and the other person will be paid a bonus according to your proposed split. Once you make your decision, the game is over and you and the Role B person will have no other opportunities to affect each other's bonuses.

Please answer these practice questions below.

If you offer 5 out of 30 cents to the Role B person, what bonus will the Role B person get (in cents)?

- 15 cents
- 5 cents
- 10 cents

If you offer 20 out of 30 cents to the Role B person, what bonus will you get (in cents)?

- 30 cents
- 20 cents
- 10 cents

>>

Before you make your decision, you will have to answer the following questions to make sure you understand the rules.
If you answer any of the questions incorrectly, you will NOT earn a bonus.

Which decision by Player A (You) makes both players earn the same bonus?

- Give Player B none of the bonus
- Give Player B half of the bonus
- Give Player B the entire bonus

Which decision by Player A (You) makes you earn the largest bonus?

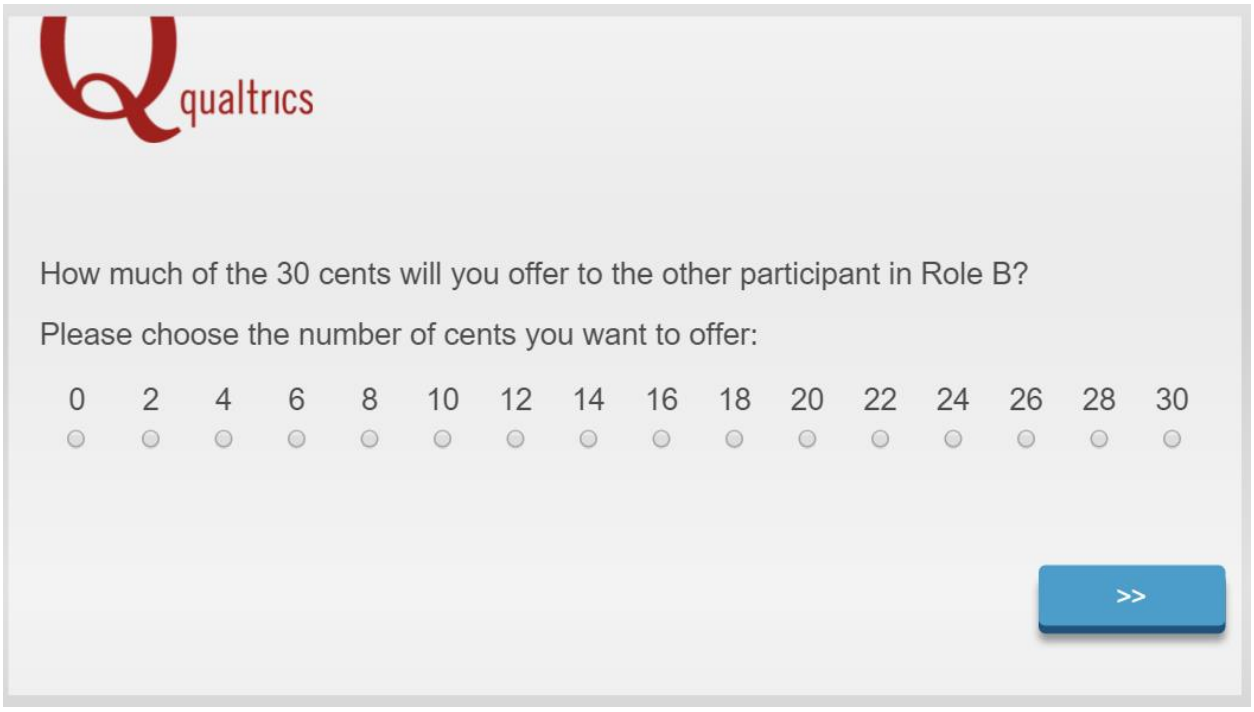
- Give Player B none of the bonus
- Give Player B half of the bonus
- Give Player B the entire bonus

Which decision by Player A (You) makes Player B earn the largest bonus?

- Give Player B none of the bonus
- Give Player B half of the bonus
- Give Player B the entire bonus

>>


qualtrics
You will now continue on to the actual real decision. Keep in mind the practice you just went through as you go on to the next section.
>>


qualtrics
How much of the 30 cents will you offer to the other participant in Role B?
Please choose the number of cents you want to offer:
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
>>

In this part of the HIT, you will play in a three-person game. You have been randomly assigned to interact with two other MTurk workers. You have been assigned to be Player 3. The other people will be Players 1 and 2.

All three of you will receive this same set of instructions. Once you make your decision you will have no other interaction with these players.
You can only play this game once.

In addition to the payment you each receive for participating in this HIT, you can earn more as a **bonus**, as follows:

In Stage 1:

Players 1 and 2 each start with 15 cents.

Player 1 makes a choice to either:

- Give their 15 cents to Player 2 (so that Player 1 gets 0 and Player 2 gets 30 cents)
- Do nothing (so both players get 15 cents)
- Take 15 cents from Player 2 (so Player 1 gets 30 and Player 2 gets 0).

Player 2 can do nothing and must accept Player 1's decision.

In Stage 2:

Player 3 also starts with 15 cents, and finds out about Player 1's choice in Stage 1.

Player 3 can spend up to their full 15 cents to reduce Player 1's bonus. For every cent Player 3 spends, Player 1 loses 2 cents.

Summary:

Therefore, Player 1's total bonus is

- The money they decide to keep
- Plus the money they decide to take from Player 2
- Minus the money Player 3 deducts from them

Player 2's total bonus is
- The money they start with
- Minus the money Player 1 takes from them

Player 3's total bonus is
- The mone they start with
- Minus the money they spend on reducing Player 1's bonus

Note: If Player 1's earnings in Stage 1 are less than the amount they are punished by Player 3, they will not go negative but will just have their bonus brought to zero.

Please answer these practice questions below.

Imagine that Player 1 is deciding whether to share, do nothing or take from Player 2.

If Player 3 decides not to reduce Player 1's bonus, which decision will result in **Player 1** earning the full 30 cents?

- ○ Player 1 deciding to share
- ○ Player 1 deciding to take
- ○ Player 1 deciding to do nothing in round 1

Imagine that Player 3 is deciding whether or not to reduce Player 1's bonus.

Which punishment amount would Player 3 chose to reduce **Player 1**'s earning down to **0 cents**, if Player 1 enters round 2 with 15 cents?

- ○ Player 3 deciding to punish 5 cents
- ○ Player 3 deciding to punish 7 cents
- ○ Player 3 deciding to punish 10 cents

Imagine that Player 1 is deciding whether to share, do nothing or take from Player 2.

Which decision will result in **Player 2** earning 0 cents?

- ○ Player 1 deciding to take

- Player 1 deciding to share
- Player 1 deciding to take
- Player 1 deciding to do nothing in round 1

Imagine that Player 3 is deciding whether or not to reduce Player 1's bonus.

Which punishment amount would Player 3 chose to reduce **Player 1**'s earning down to **0 cents**, if Player 1 enters round 2 with 15 cents?

- Player 3 deciding to punish 5 cents
- Player 3 deciding to punish 7 cents
- Player 3 deciding to punish 10 cents

Imagine that Player 1 is deciding whether to share, do nothing or take from Player 2.

Which decision will result in **Player 2** earning 0 cents?

- Player 1 deciding to take
- Player 1 deciding to do nothing in round 1
- Player 1 deciding to share

Imagine that Player 3 is deciding whether or not to reduce Player 1's bonus.

If **Player 3** decides to reduce Player 1's earning down to **10 cents**, how much would he have left after paying to reduce Player 1's amount if Player 1 enters round 2 with 30 cents?

- Player 3 would have 5 cents left
- Player 3 would have 10 cents left
- Player 3 would have 15 cents left

>>

Before you make your decision, you will have to answer the following questions to make sure you understand the rules.
If you answer any of the questions incorrectly, you will NOT earn a bonus.

Imagine that Player 1 is deciding whether to share, do nothing or take from Player 2.

If Player 3 decides not to reduce Player 1's bonus, which decision will result in **Player 1** earning the highest payoff?

- Player 1 deciding to share
- Player 1 deciding to take
- Player 1 deciding to do nothing

Imagine that Player 1 is deciding whether to share, do nothing or take from Player 2.

Which decision will result in **Player 2** earning the highest payoff?

- Player 1 deciding to share
- Player 1 deciding to take
- Player 1 deciding to do nothing

Imagine that Player 3 is deciding whether or not to reduce Player 1's bonus.

Which decision will result in **Player 1** earning the highest payoff?

- Player 3 deciding to reduce Player 1's bonus
- Player 3 deciding NOT to reduce Player 1's bonus
- Neither - Player 1's payoff is not influenced by 3's decision

If Player 3 decides not to reduce Player 1's bonus, which decision will result in **Player 1** earning the highest payoff?

- Player 1 deciding to share
- Player 1 deciding to take
- Player 1 deciding to do nothing

Imagine that Player 1 is deciding whether to share, do nothing or take from Player 2.

Which decision will result in **Player 2** earning the highest payoff?

- Player 1 deciding to share
- Player 1 deciding to take
- Player 1 deciding to do nothing

Imagine that Player 3 is deciding whether or not to reduce Player 1's bonus.

Which decision will result in **Player 1** earning the highest payoff?

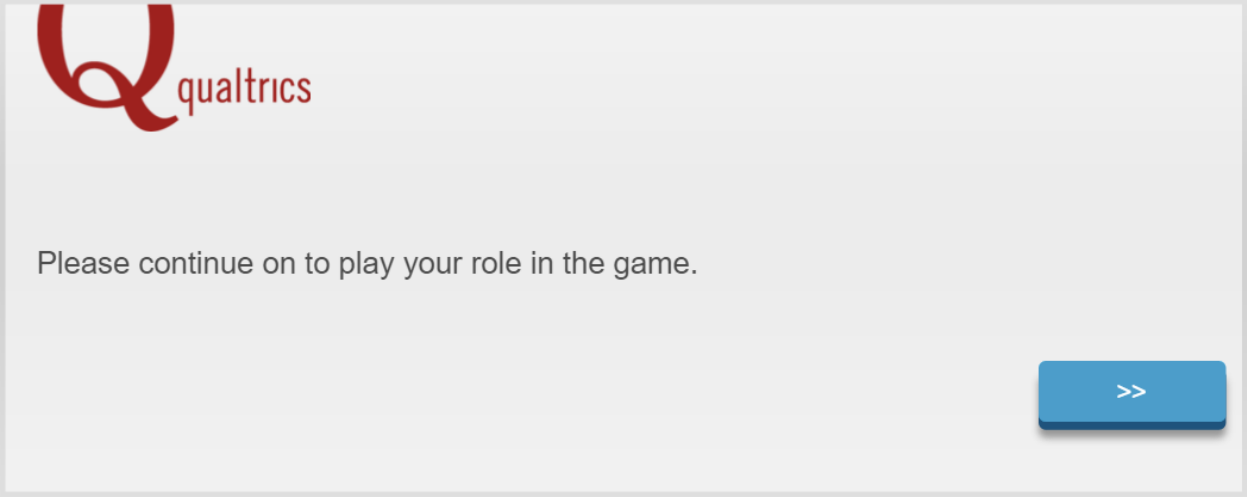
- Player 3 deciding to reduce Player 1's bonus
- Player 3 deciding NOT to reduce Player 1's bonus
- Neither - Player 1's payoff is not influenced by 3's decision

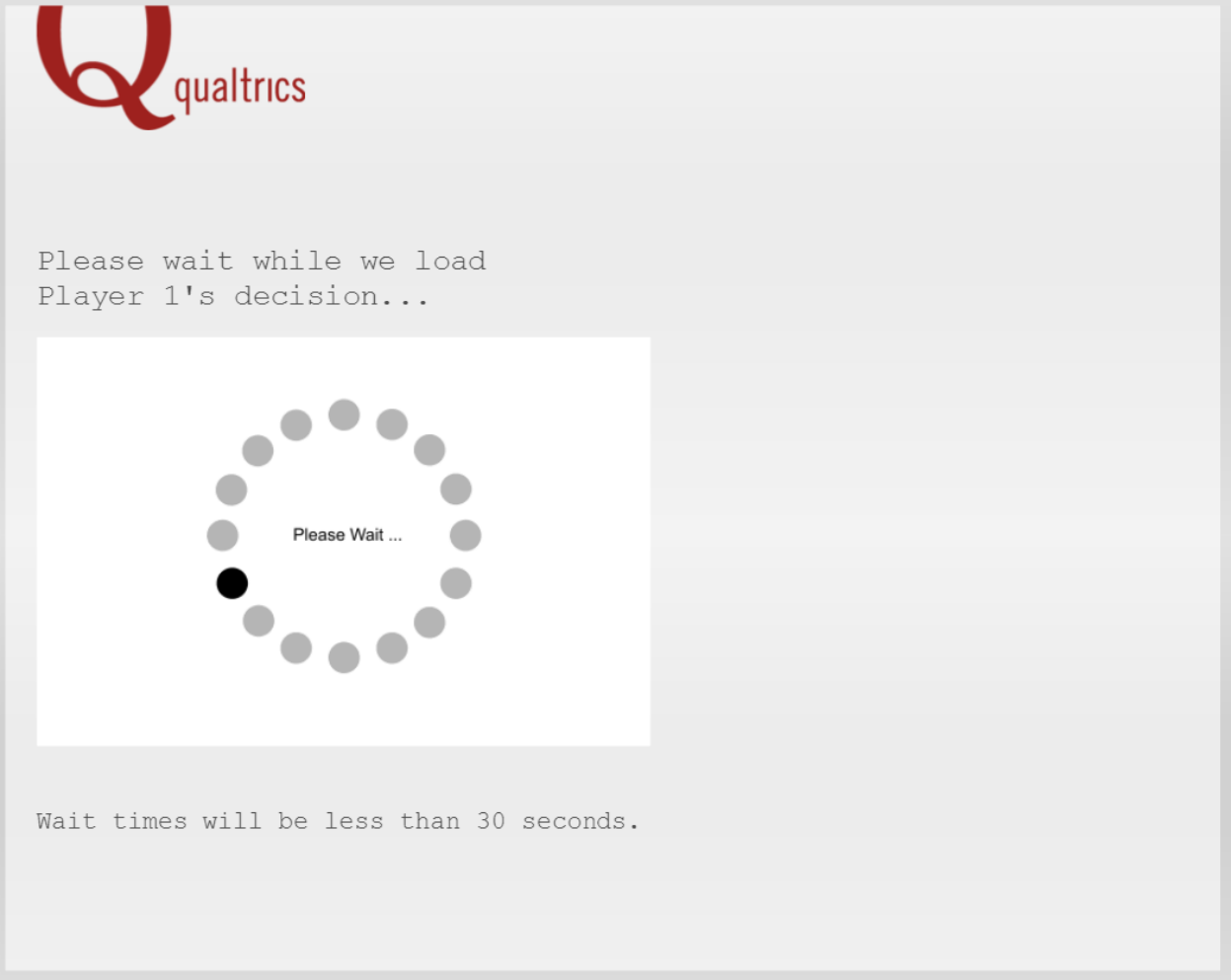
Imagine that Player 3 is deciding whether or not to reduce Player 1's bonus.

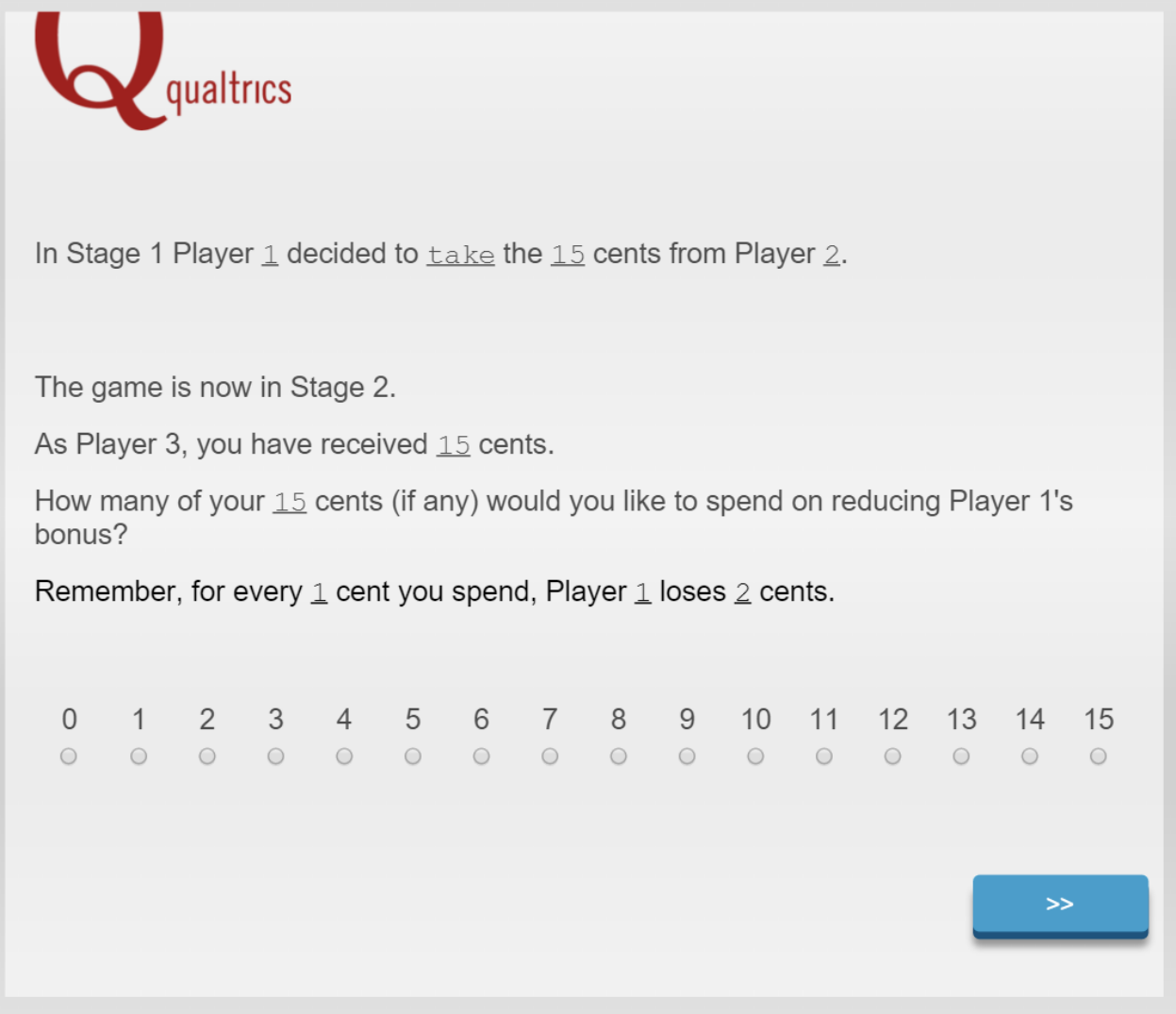
Which decision will result in **Player 3** earning the highest payoff?

- Player 3 deciding to reduce Player 1's bonus
- Player 3 deciding NOT to reduce Player 1's bonus
- Neither - Player 3's payoff is not influenced by 3's punishing decision

>>


qualtrics
Please continue on to play your role in the game.
>>


qualtrics
Please wait while we load
Player 1's decision...
Please Wait ...
Wait times will be less than 30 seconds.


qualtrics
In Stage 1 Player 1 decided to take the 15 cents from Player 2.
The game is now in Stage 2.
As Player 3, you have received 15 cents.
How many of your 15 cents (if any) would you like to spend on reducing Player 1's bonus?
Remember, for every 1 cent you spend, Player 1 loses 2 cents.
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
>>

Please answer the following questions about yourself as accurately as possible.

Your gender:

- Male
- Female

Your age:

What is the highest level of education you completed:

- Less than a high school degree
- High School Diploma
- Vocational Training
- Attended College
- Bachelor's Degree
- Graduate Degree
- Unknown

What is your religious affiliation?

- Christian
- Muslim
- Jewish
- Hindu
- Buddhist

- Atheist
- Agnostic
- Other

What most closely represents your ethnicity?

- Caucasian
- African American
- East Asian
- Pacific Islander
- Latino/ Latina
- Mid Eastern
- Indian
- Native American
- Other

Politically, how conservative are you in terms of social issues

1 - Very conservative	2	3	4	5	6	7 - Very liberal
○	○	○	○	○	○	○

Politically, how conservative are you in terms of fiscal issues

1 - Very conservative	2	3	4	5	6	7 - Very liberal
○	○	○	○	○	○	○

Please choose the category that describes the total amount of income you earned in 2013. Consider all forms of income, including salaries, tips, interest and dividend payments, scholarship support, student loans, parental support, social security, alimony, and child support, and others.

- Other

Politically, how conservative are you in terms of social issues

1 - Very conservative	2	3	4	5	6	7 - Very liberal
○	○	○	○	○	○	○

Politically, how conservative are you in terms of fiscal issues

1 - Very conservative	2	3	4	5	6	7 - Very liberal
○	○	○	○	○	○	○

Please choose the category that describes the total amount of income you earned in 2013. Consider all forms of income, including salaries, tips, interest and dividend payments, scholarship support, student loans, parental support, social security, alimony, and child support, and others.

- Under $5,000
- $5,000-$10,000
- $10,001-$15,000
- $15,001-$25,000
- $25,001-$35,000
- $35,001-$50,000
- $50,001-$65,000
- $65,001-$80,000
- $80,001-$100,000
- Over $100,000

>>

We are going to name a number of organizations. For each one, could you tell us how much confidence you have in them.
Is it:
- A great deal of confidence
- Quite a lot of confidence
- Not very much confidence
- None at all

How much confidence do you have in the Police?

- A great deal of confidence.
- Quite a lot of confidence.
- Not very much confidence.
- None at all.

How much confidence do you have in the Courts?

- A great deal of confidence.
- Quite a lot of confidence.
- Not very much confidence.
- None at all.

How much confidence do you have in your Government (in your nation's capitol)?

- A great deal of confidence.
- Quite a lot of confidence.
- Not very much confidence.
- None at all.

How much confidence do you have in your Government (in your nation's capitol)?

- A great deal of confidence.
- Quite a lot of confidence.
- Not very much confidence.
- None at all.

How much confidence do you have in Political Parties?

- A great deal of confidence.
- Quite a lot of confidence.
- Not very much confidence.
- None at all.

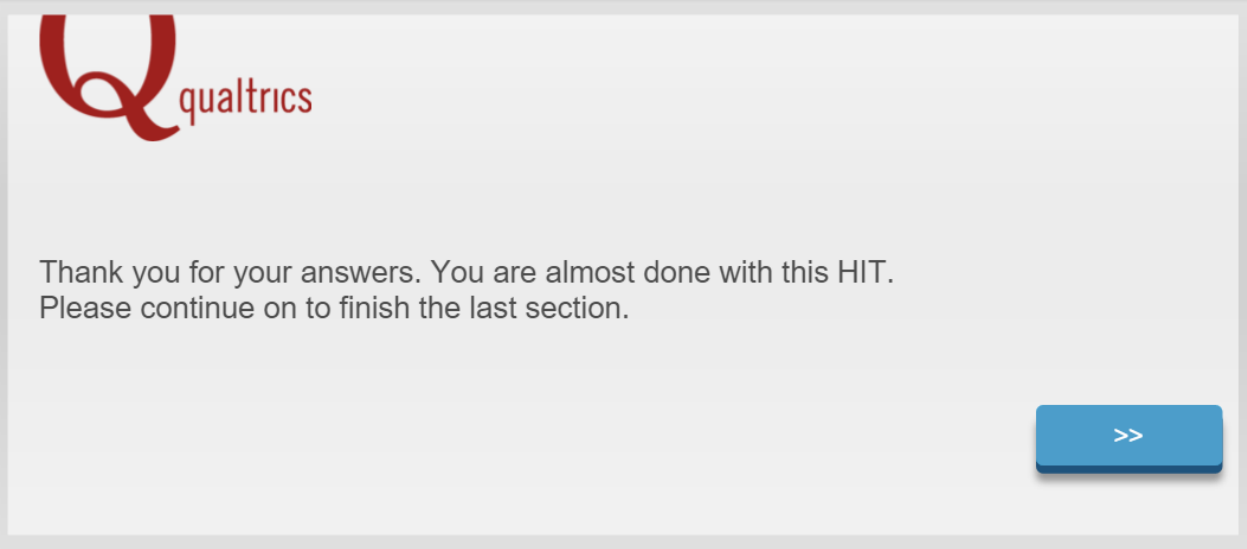
How much confidence do you have in Civil Services?

- A great deal of confidence.
- Quite a lot of confidence.
- Not very much confidence.
- None at all.

How much confidence do you have in the Banks?

- A great deal of confidence.
- Quite a lot of confidence.
- Not very much confidence.
- None at all.

>>


qualtrics
Thank you for your answers. You are almost done with this HIT.
Please continue on to finish the last section.
>>

People sometimes describe themselves as belonging to the working class, the middle class, or the upper or lower class. Would you describe yourself as belonging to the:

Low Class	Working Class	Low Middle Class	Upper Middle Class	Upper Class
○	○	○	○	○

In this question there is an income scale on which 1 indicates the lowest income group and 10 the highest income group in your country. We would like to know in what group your household is. Please, specify the appropriate number, counting all wages, salaries, pensions and other incomes that come in.

Lower Step	Second Step	Third Step	Fourth Step	Fifth Step	Sixth Step	Seventh Step	Eighth Step	Ninth Step	Tenth Step
○	○	○	○	○	○	○	○	○	○

What is your profession?

To what extent do you feel you can trust other people that you interact with in your daily life?

1 - Very little	2	3	4	5	6	7 - Very much
○	○	○	○	○	○	○

I trust my initial feelings about people.

1 - Very untrue	2	3	4	5 - Very true

To what extent do you feel you can trust other people that you interact with in your daily life?

1 - Very little	2	3	4	5	6	7 - Very much
○	○	○	○	○	○	○

I trust my initial feelings about people.

1 - Very untrue	2	3	4	5 - Very true
○	○	○	○	○

I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.

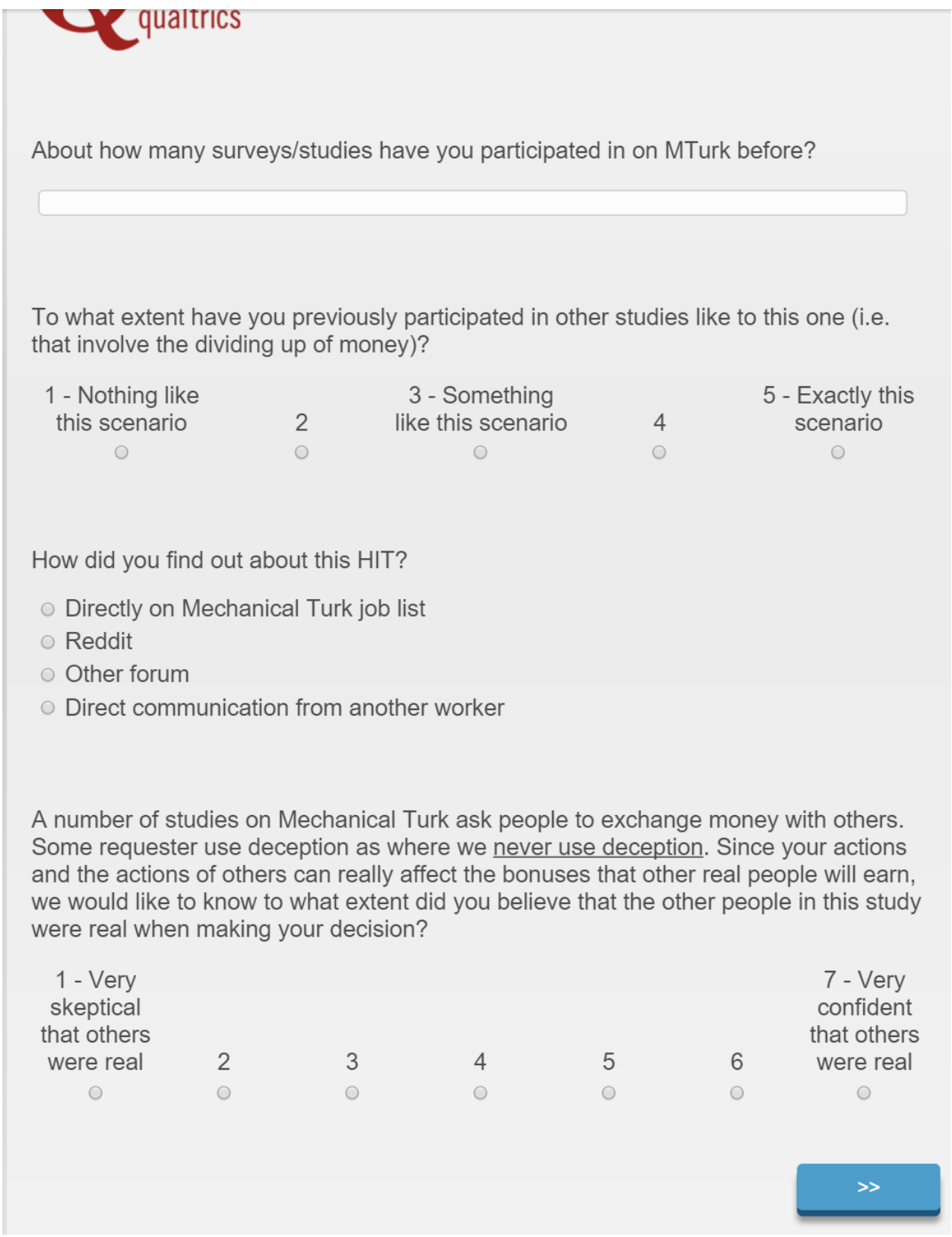
1 - Very untrue	2	3	4	5 - Very true
○	○	○	○	○

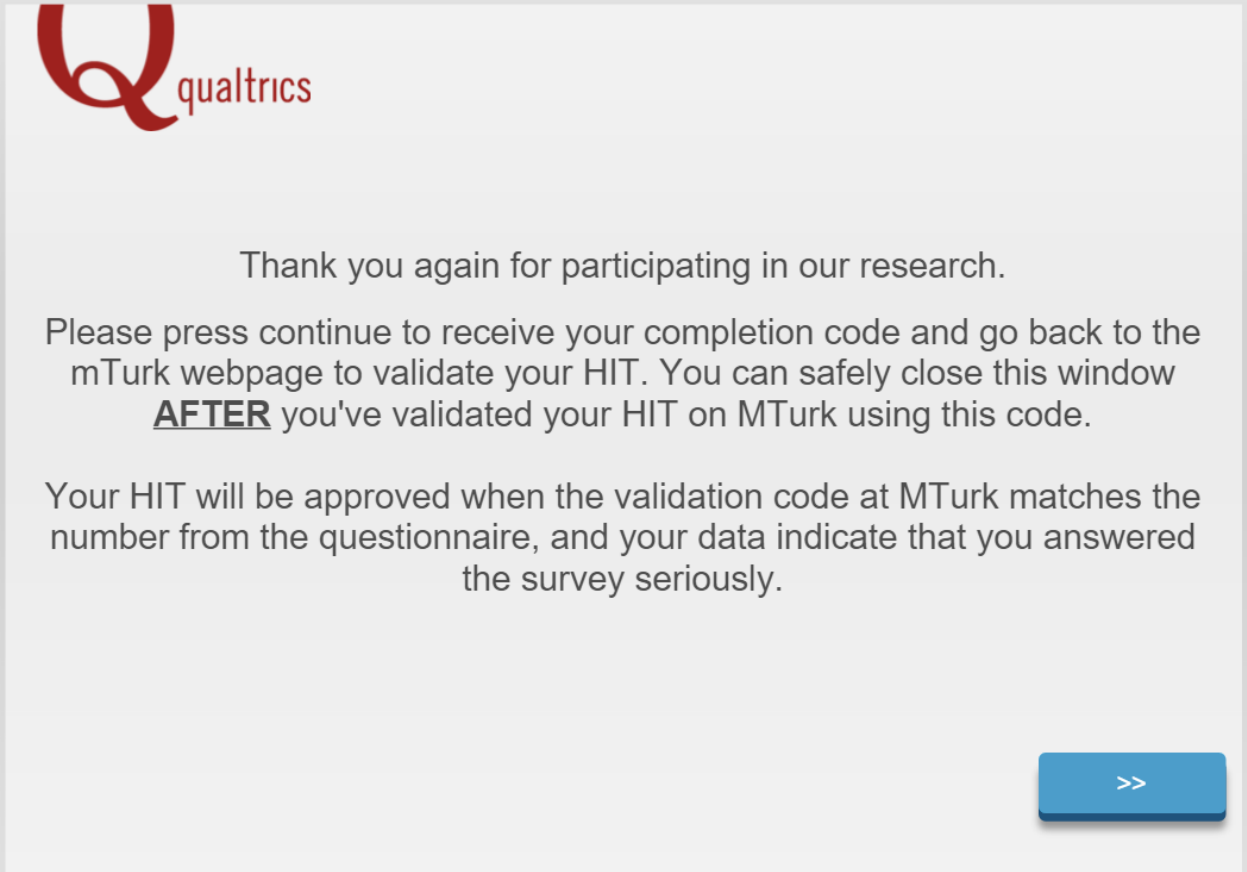
Please take a second and explain why you made the decisions you did in each game.

When you were Player A giving to Player B:

When you were Player 3 punishing Player 1:

>>


qualtrics
About how many surveys/studies have you participated in on MTurk before?
To what extent have you previously participated in other studies like to this one (i.e. that involve the dividing up of money)?
1 - Nothing like this scenario
2
3 - Something like this scenario
4
5 - Exactly this scenario
How did you find out about this HIT?
Directly on Mechanical Turk job list
Reddit
Other forum
Direct communication from another worker
A number of studies on Mechanical Turk ask people to exchange money with others. Some requester use deception as where we never use deception. Since your actions and the actions of others can really affect the bonuses that other real people will earn, we would like to know to what extent did you believe that the other people in this study were real when making your decision?
1 - Very skeptical that others were real
2
3
4
5
6
7 - Very confident that others were real
>>


qualtrics
Thank you again for participating in our research.
Please press continue to receive your completion code and go back to the mTurk webpage to validate your HIT. You can safely close this window AFTER you've validated your HIT on MTurk using this code.
Your HIT will be approved when the validation code at MTurk matches the number from the questionnaire, and your data indicate that you answered the survey seriously.
>>

Study 2

Instructions

The setup of this HIT is **different** from other tasks that you might be used to completing on MTurk. Here you will be playing with **real people** that also accepted this HIT, who are completing it **at the same time**. It is therefore very important that you complete this task **without interruptions**. Including the time for reading these instructions, the HIT will take about 15 minutes to complete.

When you confirm having read these instructions, the task itself will be started in a pop-up window. During the task, please **do not close this window** or leave the task's web pages in any other way. If you do close your browser or leave the task, you will not be able to re-enter the task and we will not be able to pay you!

This task consists of **3 parts** where you can earn **points**. At the end of the HIT **only one part** will be **randomly** selected and paid to you at the following exchange rate: **1 point = $0.001**. If your total in the selected Part is negative you get no bonus.

You must therefore pay attention to the three parts. Further information about each part will be provided later.

It is very important that you make your decision within the **time limit** shown on your screen. If you fail to do so, you will be **excluded from the task and not receive any payment**.

Please click the link below if you understood the instructions.

Before you do so, please remember the following:

- The task consists of 3 parts and **only one** will be paid.
- You will make decisions with **real people**.
- All players complete this task **at the same time**.
- So please complete this task **without interruptions**.
- The task will open in a pop-up window. Please allow for pop-ups in your browser!

Do not close the pop-up window during the task because this will terminate the task and we will not be able to pay you!

I have read and understood the instructions. Continue!

Part 1

In this first part you will play a game **with 2 people** for **10 rounds**. In each round each of you will receive 140 points and have to decide **how many points** to contribute to a group project.

After your decisions, all points contributed to the **group project** are added up, and this number of points is **multiplied by 1.2**. The resulting number of points is then divided equally between you and the other two participants (irrespective of how much each person contributed).

Thus for every 10 points you contribute, you get only 4 points back: so no matter what the other group members contribute, you personally lose points on contributing, but contributing benefits the group as a whole.

For example:
* If everyone contributes their 140 points, everyone's points will increase by 1.2: each of you will earn 168 points.
* If everyone else contributes their 140 points, while you keep your 140 points, you will earn 252 points, while the others will earn only 112 points.

An Inspection Mechanism

There is a **20 in 100** chance of having your contribution **inspected** in each round.
If you are inspected, you will be **fined** 1.5 points for every 1 point you chose to keep for yourself in this round.
Therefore, anyone who contributes less than the maximum of 140 points will be fined **if inspected.**

For example:
* If you contributed 50 points, and then got inspected, you would be fined 135 points (Contributing 50 points means you kept 90 out of the 140 points for yourself. Therefore you would be fined 90 x 1.5 = 135 points.).

Before beginning the task, please answer these questions to make sure you understand the rules.

1. How much would you contribute to earn the most points for the group as a whole?
I would contribute _____ points.

2. If you do not get inspected, what contribution amount earns the most points for you personally?
When I contribute _____ points.

3. If you do get inspected, what contribution amount earns the most points for you personally?
When I contribute _____ points.

Submit

Thank you for your patience!

Please wait until the other participants have read the instructions and completed the quiz.

Note that this could take a few minutes.

We are waiting for **2** more players...

As soon as the other participants are ready, the task will start automatically.
From that point onwards, the task should take approximately 10 minutes to complete.

If you are still waiting when the time below is up,
you will be asked if you want to leave this task.
If you choose to leave, you will receive a code to directly collect your participation fee for this HIT.

Remaining time: 02:00

All players are here. The task starts now!

Remaining time: 34

Part 1

Round 1

A new round has started.
You received **140** points.

How many points do you want to contribute to the project?

OK

Remaining time: 0

Part 1

Round 1 Contributions

You	Others (in random order)	
140	60	100

Round 1 Earnings

Points kept for yourself	0
Points received from the group project	150
You were inspected	0
Your total in this round	**150**

Your total including this round	**150**

Part 2

In this part you will play a game **with another two people** for **this round only**. Each of you start with **1400 points**.

The game consists of **two stages**.

In the first stage, Player 1 makes a choice to either:

- **Give** 1400 points to Player 2 (so that Player 1 gets **0 points** and Player 2 gets **2800 points**)
- **Do nothing** (so both players get **1400 points**)
- **Take** 1400 points from Player 2 (so Player 1 gets **2800 points** and Player 2 gets **0 points**)

Player 2 can do nothing and must accept Player 1's decision.

In the second stage, Player 3 can spend up to their full **1400 points** to reduce Player 1's bonus. For every point Player 3 spends, Player 1 loses **2** points.

Summary:

Player 1 earns: 1400 points ***plus / minus*** the points *taken from / given to* Player 2 ***minus*** the points Player 3 deducts from them

Player 2 earns: 1400 points ***plus / minus*** the points Player 1 *takes from / gives to* them

Player 3 earns: 1400 points ***minus*** the points they spend on reducing Player 1's bonus

Before beginning the task, please answer these questions to make sure you understand the rules.

1. When does Player 1 ***earn*** *more?*
When Player 1 decides to take the points and Player 3 spends [ ] points.

2. When does Player 1 ***lose*** *more?*
When Player 1 decides to take the points and Player 3 spends [ ] points.

Submit

You have been assigned to be Player 3.

Please wait

Remaining time: 45

Part 2

The game is now on Stage 2. In the previous stage Player 1 decided to take the points.
Therefore, **Player 1 got 2800 points, player 2 got 0 points**.

As player 3, you have received **1400** points.
How many of your **1400** points (if any) would you like to spend on reducing Player 1's bonus?

Remember, for every point you spend, Player 1 loses 2 points.

OK

Part 3

In this part you will play a game **with another new person** for **this round only**.

You have been assigned to be in Role **A** and the other person will be in Role **B**.

You will get to make an offer of how to split **1400 points** between you and the person in Role B. The person in Role B must accept this offer, so you and the other person will be paid a bonus according to your proposed split. Once you make your decision, the game is over and you and the Role B person will have no other opportunities to affect each other's bonuses.

Before beginning the task, please answer these questions to make sure you understand the rules.

1.As the person in Role A, when do you personally earn more?
When I make an offer of points.

1.As the person in Role A, when does the person in Role B earn more?
When I make an offer of points.

Submit

Remaining time: 26

Part 3

You received **1400** points.

How many points will you send to the other person? [ ]

OK

Summary

Thank you very much for your participation!

This task consisted of 3 parts and one of them will be selected randomly to be paid in bonus.

In Part 1 you received 150 points; in Part 2 you received 1300 points and in Part 3 you received 700 points.

The part selected for payment is Part 3. Therefore, you will get a bonus of $0.70.

To receive your payment please copy the following unique code and paste it into MTurk:

93751553

Please note that you will first receive a payment with your show-up fee and within 24 hours another payment with the bonus here presented.

After you have entered your code, you can close this window.

Appendix C

Here we complement the main text analyses, which examine the relationship between institutional quality and average amount of DG giving/TPPG punishment, with analyses that separately consider the probability of giving/punishing and the average amount of DG giving/TPPG punishment among those who give/punish.

Study 1

With respect to *probability* of giving in the DG (Figure C1a), logistic regression revealed a significant positive effect of institutional quality, b= 0.346, z= 2.42, p= 0.016; including controls, b= 0.307, z= 1.93, p= 0.053. No significant interaction was observed between institutional quality and game comprehension, b= 0.971, z= 1.64, p= 0.102, and excluding those participants who failed game comprehension from our main analysis produced no qualitative difference in the relationship between institutional quality and probability of giving, b= 0.293, z= 1.96, p= 0.050.

With respect to amount given conditional on giving (Figure C1b), linear regression found no significant effect of institutional quality either excluding controls β= 0.011, t= 0.27, p= 0.790, or when including controls, β= 0.025, t= 0.55, p= 0.582; and there was no significant interaction between institutional quality and game comprehension, β= -0.007, t= 0.05, p= 0.964 or between institutional quality and stage order (p= 0.689). However, a trending interaction between institutional quality and game order was observed (β= 0.379, t= 1.74, p= .083), such that a trending negative relationship between institutional quality and amount given emerged only in the [TPPG,DG] order, β= -0.113, t= -1.73, p= 0.085, but not in the [DG,TPPG] order, β= 0.038, t= 0.67, p= 0.504.

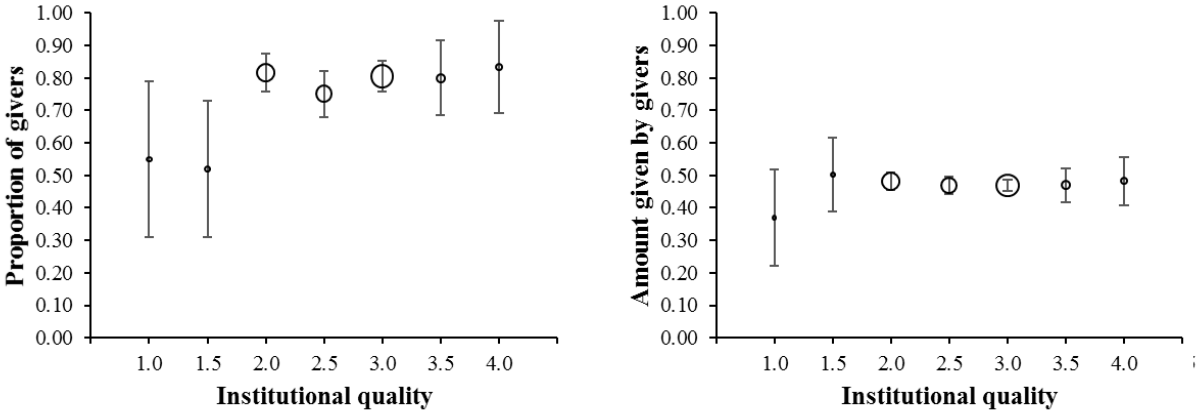


Fig. C1. *Effect of daily life institutional quality on the amount given in the (a) probability of giving in the DG, and (b) amount given in the DG among those who gave a non-zero amount in Study 1 (shown as fraction of the Dictator's endowment). Institutional quality was binned by rounding up to the nearest 0.5, and average value across all participants in each bin is shown; dot size is proportional to the number of observations in each bin. Error bars indicate 95% confidence intervals.*

Next we examined punishment in the TPPG. For *probability* of punishing (Figure C2a), logistic regression did not find a significant effect of institutional quality, b= 0.156, z= 1.29, p= 0.198. This relationship remained non-significant after including controls, b= 0.161, z= 1.18, p= 0.238; no significant interaction was observed between institutional quality and game comprehension, b= -0.039, z= 0.15, p= 0.884; and there were no significant interactions between institutional quality and order (p>.58 for all).

With respect to *amount* of punishment conditional on punishing (Figure C2b), linear regression found no significant effect of institutional quality, either excluding controls, β= -0.083, t= -1.39, p= 0.166, or including them, β= -0.537, t= -1.16, p= 0.245. There was, however, a significant interaction between institutional quality and game comprehension, β= 0.65, t= 2.35 p= 0.019, such that there was an unanticipated significant *negative* relationship between institutional quality and amount of punishment among perfect comprehenders, β= -0.22, t= -2.37, p= 0.019, but a non-significant positive relationship among participants who answered one or more TPPG questions incorrectly, β= 0.0.322, t= 0.37, p= 0.714. There was no significant interaction between institutional quality and stage order (p= 0.674). There was a significant interaction between institutional quality and game order (β= 0.796, t= 2.63, p= 0.009), however decomposing this interaction showed no significant simple effects of institutional quality on amount punished in other order (p>0.19 for both).

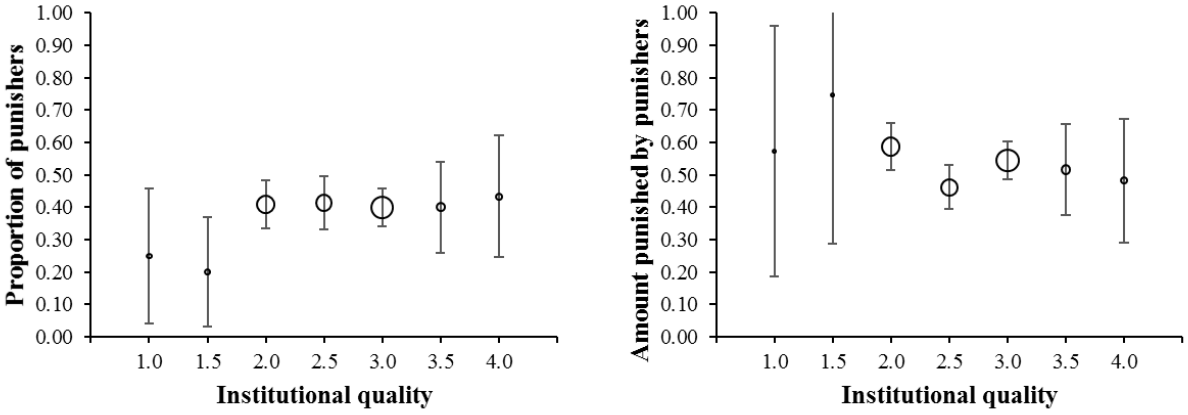


Fig. C2. Effect of daily life institutional quality on (a) probability of punishing in the TPPG, and (b) amount spent on punishing in the TPPG among those who punished a non-zero amount in Study 1 (shown as fraction of the Sanctioner's endowment). Institutional quality was binned by rounding to the nearest 0.5, and average value across all participants in each bin is shown; dot size is proportional to the number of observations in each bin. Error bars indicate 95% confidence intervals.

Study 1R

As in study 1, we find that institutional quality has a significant positive effect on the probability of giving (Figure C2a), b= 0.352, z= 3.76, p< 0.001, which is robust to the inclusion of demographic controls, b= 0.293, z= 2.76, p= 0.006. A significant *positive* interaction was observed between institutional quality and game comprehension, b= 0.713, z= 2.03, p= 0.043, but excluding those participants who failed game comprehension from our main analysis produced no qualitative difference in the relationship between institutional quality and probability of giving, b= 0.288, z= 2.93, p= 0.003.

We also find no significant effect of institutional quality and amount given, conditional on giving (Figure C1b); either excluding controls β= 0.002, t= 0.06, p= 0.949, or when including controls, β= -0.012, t= 0.41, p= 0.679. A significant positive interaction between institutional quality and game comprehension was also observed, β= 0.256, t= 2.17, p= 0.030, but excluding those participants who failed game comprehension from our main analysis produced no qualitative difference in the relationship studied, β= -0.094, t= -0.62, p= 0.536.

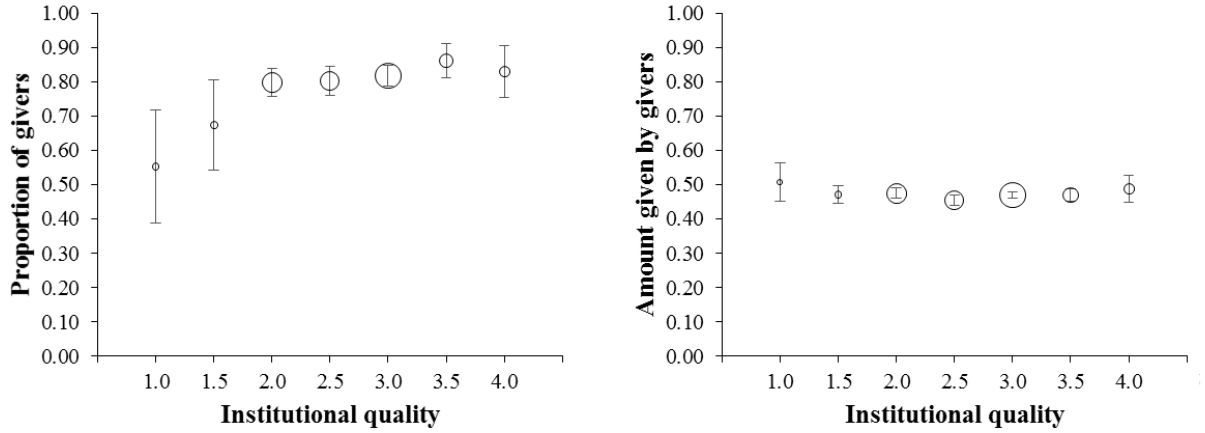


Fig. C3. *Effect of daily life institutional quality on the amount given in the (a) probability of giving in the DG, and (b) amount given in the DG among those who gave a non-zero amount in Study 1R (shown as fraction of the Dictator's endowment). Institutional quality was binned by rounding up to the nearest 0.5, and average value across all participants in each bin is shown; dot size is proportional to the number of observations in each bin. Error bars indicate 95% confidence intervals.*

Study 2

With respect to *probability* of giving in the DG (Figure C3a), participants were substantially more likely to give in the high institutional quality condition (69% givers) compared to the low institutional quality condition (55% givers) (logistic regression, b= 0.596, z= 2.59, p= 0.01). Institutional quality did not interact significantly with game order, b= -0.528, z= -1.14, p=

0.253 (although we note that, reassuringly, in the "cleaner" order in which the DG was played first the effect of institutional quality was larger than overall: low, 58% givers; high, 78% givers).

With respect to *amount* given in the DG conditional on giving, linear regression found no significant effect of institutional quality on amount given, β= 0.021, t= 0.41, p= 0.681 (Figure C3b). Among participants who gave in the DG, a vast majority gave away exactly half of the endowment, for both the low institutional quality (96% of givers) and high institutional quality (94% of givers). Institutional quality interacted only marginally with game order, β= -0.215, t= -1.67, p= 0.097. Decomposing this marginally significant interaction with game order showed no significant effect of institutional quality in either order, *ps*> 0.097.

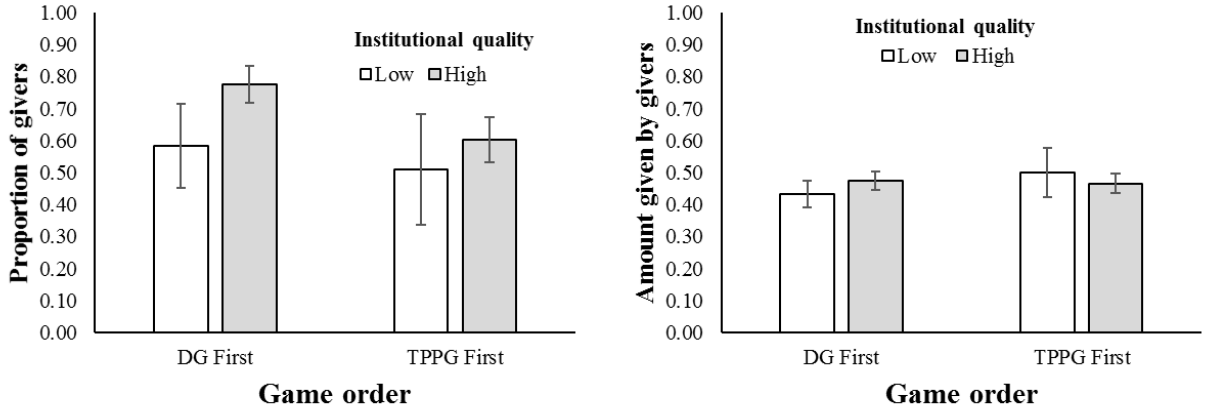


Fig. C4. Effect of experimentally induced institutional quality on (a) probability of giving in the DG, and (b) amount given in the DG among those who gave a non-zero amount (shown as fraction of the Dictator's endowment) in Study 2. Error bars indicate 95% confidence intervals.

With respect to *probability* of punishing (Figure C4a), in contrast to Study 1's correlational result, logistic regression found no significant effect of institutional quality on punishing, b= 0.285, z= 1.31, p= 0.19. There was a marginally significant interaction between institutional quality and game order, b= -0.818, z= -1.9, p= 0.057. Decomposing this interaction found that when the DG came before the TPPG, institutional quality led to a significantly higher likelihood of choosing to punish, b= 0.68, z= 2.14, p= 0.033, whereas in the cleaner order where the TPPG came before the DG, there was no significant effect of institutional quality b= -0.139, z= -0.48, p= 0.634. Thus it seems that institutional quality had little direct effect on probability of punishing, and to whatever extent there was an effect, it only occurred when participants had made their own prosocial decision as a Dictator immediately before having to sanction someone else for failing to be prosocial.

With respect to *amount* of punishment conditional on punishing (Figure C4b), linear regression found no significant relationship between institutional quality and punishment, β= -

0.016, t= -0.24, p= 0.808, as well as no significant interaction with game order, β= -0.103, t= -0.63, p= 0.529. There was, however, a significant interaction between institutional quality and comprehension, β= -0.526, t= -2.15, p= 0.034, such that there was a non-significant negative effect of institutional quality among participants who got all the TPPG comprehension questions correct on the first try, β= -0.101, t= -1.41, p= 0.163, but a marginally significant positive relationship among participants who took more than one try to answer correctly, β= 0.280, t=1.73, p= 0.091. Thus we did not find substantial support for the negative correlation between institutional quality and amount of punishment conditional on punishing observed in some subsets of the data in Study 1; we therefore concluded that the Study 1 result was likely spurious.

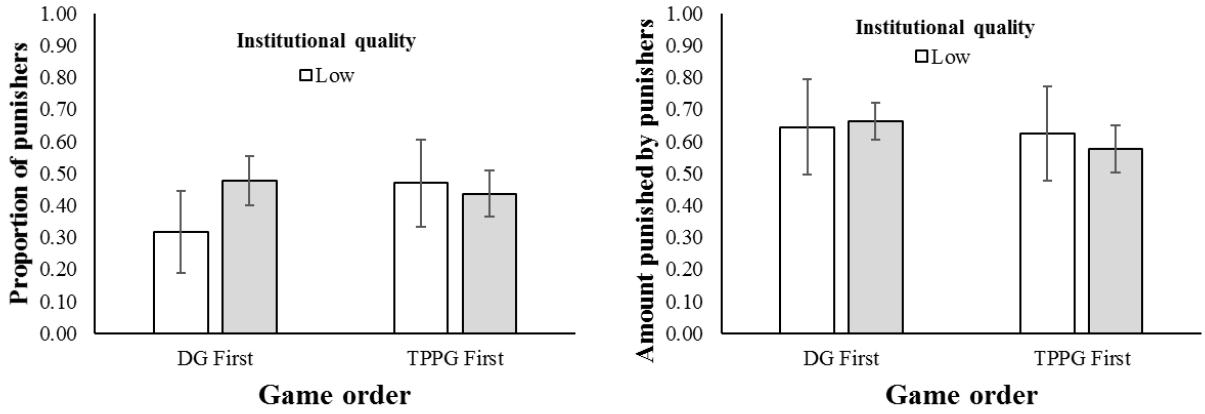


Fig. C5. Effect of experimentally induced institutional quality on (a) probability of punishing in the TPPG, and (b) amount spent on punishment in the TPPG among those who punished a non-zero amount (shown as fraction of the Sanctioner's endowment) in Study 2. Error bars indicate 95% confidence intervals.

Appendix D

Here we present the experimental design followed in Study 1R.

D.1. Materials and methods

D.1.1. Participants

We recruited 1705 MTurkers located in the USA. Participants had a mean age of 36 years and 56.5% were female. The task lasted between 5 and 10 minutes and participants received a flat payment of $0.50 for participating, plus a variable bonus that on average totaled $0.35 among those who passed the comprehension questions and thus received a bonus (min $0.05, max $0.30). We prevented repeated participation by excluding duplicate worker IDs and IP addresses within study 1R.

D.1.2. Method

This study consisted of only two stages and it was not presented in random order. Similar to the original study, in the first stage we asked participants to self-report their confidence in the police and in the courts (each rated using a four-point Likert scale ranging from "A great deal of confidence" to "None at all"). We did not include confidence in the other institutions as a precommitment to focusing only on those institutions connected with accountability. In the second stage participants played the role of the Dictator in a single-shot DG with another MTurker using neutral language. Specifically, they were asked to unilaterally choose how to divide 30 cents (in 5 cent increments) between herself and the Recipient. As in the first study, Recipients were MTurkers drawn at random from a list of worker IDs of participants in prior experiments we ran. Unlike the original study though, there was no third stage assessing participants' punishment behavior.

The experiment was performed in *Qualtrics*, and participants were not aware of the existence of subsequent stages. Specific instructions for each stage were only provided at the relevant time. Participants filled in a demographics questionnaire, similar to the one presented in the original study, upon completion of the two stages. To assess comprehension, participants were asked to complete a quiz on the rules of the DG and were told that bonuses would only be paid if answers to the quiz were correct (91.2% of subjects answered all quiz questions correctly).

Appendix E

Here we investigated whether the effect of institutional quality on DG giving varied by participant gender. As can be seen in the regression analysis below, we found no significant interaction between gender and institutional quality when predicting DG giving in either study.

Table E1
Dictator game giving in Study 1 and Study 2.

	Study 1		*Study 1R*		*Study 2*	
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Institutional quality	0.827*	1.117*	0.727**	1.184**	0.074**	0.074
	(0.416)	(0.557)	(0.238)	(0.348)	(0.039)	(0.041)
Female	1.179*	2.904	1.719***	4.029**	0.039	0.039
	(0.530)	(2.263)	(0.314)	(1.321)	(0.023)	(0.057)
Female x		-0.657		-0.857		0.001
Institutional quality		(0.838)		(0.476)		(0.062)
Constant	8.325***	7.579***	8.419***	7.200***	0.232***	0.232***
	(1.130)	(1.477)	(0.676)	(0.956)	(0.027)	(0.036)
N	707	707	1705	1705	512	512
R^2	0.014	0.014	0.024	0.025	0.019	0.019

Note. Standard errors (s.e.) and standardized betas reported; * p<0.05, ** p<0.01, *** p<0.001.