AMERICAN SOCIOLOGICAL ASSOCIATION

Social Psychology Quarterly
XX(X) 1–20

DOI: 10.1177/0190272517703750
journals.sagepub.com/home/spq
SAGE

Costly Punishment Increases Prosocial Punishment by Designated Punishers: Power and Legitimacy in Public Goods Games

Ko Kuwabara[1] and Siyu Yu[2]

Abstract

A classic problem in the literature on authority is that those with the power to enforce cooperation and proper norms of conduct can also abuse or misuse their power. The present research tested the argument that concerns about legitimacy can help regulate the use of power to punish by invoking a sense of what is morally right or socially proper for powerholders. We tested this idea in a laboratory experiment using public goods games in which one person in each group was selected to be a "designated punisher" who could give out material punishment that was either costly or costless to the punisher. Results show that costly punishment is perceived as more legitimate (proper) than costless punishment and that designated punishers engaged in more proper ("prosocial") punishment and less abusive ("antisocial") punishment when punishment was costly. These results highlight the importance of legitimacy in both motivating and regulating the enforcement of cooperation.

Keywords

legitimacy, punishment, cooperation, public goods games, power

Imagine witnessing a neighbor who fails to pick up after his dog, a colleague who leaves the coffee pot empty, or teenagers littering in the park. Part of our civic duty—as members of organizations, communities, or societies—is to help enforce basic norms of conduct. Yet, whether people actually do so depends on a number of issues, including the cost of enforcement (Is it worth my time to confront them or alert the authorities?) and people's role (Is it appropriate for me or someone else to intervene?). Indeed, the enforcement of cooperation has long puzzled social scientists, for no rational actor should voluntarily exercise punishment in the sole interest of producing public goods that would be available to punishers and non-punishers alike (Fehr

[1]INSEAD, Singapore
[2]New York University, NY, USA

Corresponding Author:
Ko Kuwabara, INSEAD, 1 Ayer Rajah Avenue, Singapore, 138676, Singapore.
Email: kokuwabara@gmail.com

and Gächter 2000; Heckathorn 1990; Yamagishi 1986). Yet, controlled experiments have amassed abundant evidence that people punish each other significantly more than is often theorized (Balliet, Mulder, and Van Lange 2011; Chaudhuri 2011), suggesting that peer sanctioning may play a critical role in the success and survival of groups, communities, and institutions (Bowles and Gintis 2011; Ostrom 1990).

Recently, scholars have begun to question the efficacy of peer sanctions. Some have noted that peer sanctioning can sustain cooperation and increase efficiency only under favorable conditions; often, peer sanctions result in retaliation or miscoordination, creating excessive punishment (Egas and Riedl 2008; Gächter, Renner, and Sefton 2008; Heckathorn 1990; Rand and Nowak 2011). Others have argued that personally punishing wrongdoers is less common outside of laboratory studies than presumed by theory (Baldassarri and Grossman 2011; Kiyonari and Barclay 2008). These concerns have led scholars to consider enforcement by *designated* punishers. By restricting the power to punish to a solitary member (e.g., leaders, authority figures, regulatory agents), designated punishment has the potential to curtail excessive punishment. Studies have shown that designated punishment is sufficient to sustain cooperation without reducing overall collective welfare (Balliet et al. 2011; O'Gorman, Henrich, and Van Vugt 2009).

Designated punishment presents at least two critical weaknesses, however. First, it is vulnerable to abuse of power. For instance, designated punishers may engage in *antisocial punishment* (sanctioning members who cooperate more than punishers; Herrmann, Thoni, and Gächter 2008) rather than *prosocial punishment* (punishing defectors). While peer punishers may engage in antisocial punishment out of retaliation, designated punishers may engage in antisocial punishment because targets cannot retaliate. Second, vesting one person with the sole responsibility of enforcement also carries the risk that the punisher may not engage in enough enforcement (Devlin-Foltz and Lim 2008).

The present research builds on the idea that concerns about legitimacy can help regulate these patterns because legitimacy derives in part from propriety, namely, personal beliefs about what is normatively desirable, appropriate, or correct (Dornbusch and Scott 1975; Hegtvedt and Johnson 2009), such as how to exercise power in ways that meet the social approval of peers and subordinates (Biggart and Hamilton 1984; Blau 1964; Zelditch and Walker 2000). Supporting this argument, our public goods experiment shows that designated punishers engage in more proper (prosocial) punishment and less improper (antisocial) punishment when punishment is costly rather than costless to punishers. This is because compared to costless punishment, costly punishment is seen as more proper, which compels punishers to use it in more fair and appropriate ways. This pattern stands in stark contrast to the standard economics of costly punishment, which suggests less punishment of any kind as the cost of punishment increases (Egas and Riedl 2008; Horne and Cutlip 2002). Overall, by integrating research on the sociology of power and legitimacy (Hegtvedt and Johnson 2009) and the behavioral economics of public goods games (Fehr and Gächter 2000), our research advances our understanding of how concerns about legitimacy can help both motivate and regulate the enforcement of cooperation.

Legitimacy

Legitimacy is a matter of critical importance for authority. Legitimacy refers to

perceptions about what others see—apparently or presumably—as proper and valid in given situations or roles (Dornbusch and Scott 1975). Propriety concerns personal beliefs about what is normatively right (e.g., what others view as fair or appropriate; Johnson, Dowd, and Ridgeway 2006; Zelditch 2001), whereas validity refers to the degree to which a person feels obliged to obey or comply with the demands of the situation (e.g., norms, rules, roles) as "matters of objective fact" (Zelditch and Walker 1984:219). Together, legitimate entities and acts are widely accepted and "taken for granted" as normative, fair, and appropriate while illegitimate ones are met with scrutiny, disapproval, and resistance (Zelditch and Walker 1984). Without legitimacy, enforcement is less effective, potentially backfiring by provoking resentment and reducing compliance (Baldassarri and Grossman 2011; Fehr and Rockenbach 2003). Without legitimacy, people with power are also more likely to alienate others and lose their support by engaging in overly aggressive or antisocial behaviors that deviate from proper norms of conduct (Fast, Halevy, and Galinsky 2012; Zelditch 2001; Zelditch and Walker 2000). Despite its obvious relevance, legitimacy has received scant attention in studies of costly punishment until recently (e.g., Baldassarri and Grossman 2011).

While validity and propriety are both important elements of legitimacy, they can vary orthogonally; something can be reasonably valid but more or less proper (Zelditch and Walker 2000), such as a political regime (Kuran 1995) or unpopular norms (Willer, Kuwabara, and Macy 2009) that people ostensibly follow without personally endorsing them. Past research (summarized in Zelditch and Walker 2000) suggests that propriety—independent of validity—can help regulate the use of power by authority. In laboratory studies simulating bureaucratic decision making, participants assigned to proper (vs. less proper) positions were more likely to use their power in ways that they believed was proper. Building on this insight, our goal in the present research is to see how concerns about propriety in particular affect patterns of punishment in public goods games. In doing so, we highlight the relevance of legitimacy to questions of fundamental importance about human cooperation and enforcement.

Enforcers can derive legitimacy from social processes like peer election (Baldassarri and Grossman 2011; Kosfeld, Okada, and Riedl 2009) or appointment to formal roles by authority (Walker, Rogers, and Zelditch 1988). Here, we examine legitimacy that derives from using costly versus costless punishment. We argue that punishment is perceived as more proper when it is costly rather than costless to oneself because enforcing cooperation at one's own expense is viewed as selfless, fair, and sincere. In collective action settings, such acts of self-sacrifice for collective welfare are often rewarded with peer approval and compliance (Barclay 2006; Willer 2009). In this view, costly punishment is a form of costly signal that helps convey prosocial motives and affirms legitimacy in one's group if used properly (Jordan et al. 2016). Conversely, punishers may risk antagonizing their peers by misusing their power, namely, using punishment improperly or using punishment that lacks apparent legitimacy (Willer et al. 2012; Xiao 2013; Xiao and Tan 2014). Costless punishment, in other words, may be perceived as improper—selfish, unfair, and antisocial—because it penalizes the target without any cost to the punisher.

Designated Punishment

We further argue that making punishment costless rather than costly has

markedly different effects on peer versus designated punishment. The reason is that the cost of punishment helps offset concerns about the unequal distribution of power to punish under conditions of designated punishment. In peer punishment, all members are given the same role and the power to punish each other, which diffuses concerns about the fairness and appropriateness of punishment (Kurzban, DeScioli, and O'Brien 2007; Molenmaker, de Kwaadsteniet, and van Dijk 2016). Instead, under the "default" case of peer punishment, punishment is driven primarily by the basic economics of punishment, punishing less as the cost to oneself increases and the cost to targets decreases (Egas and Riedl 2008; Horne and Cutlip 2002). For instance, Egas and Riedl (2008) find that in public goods games with punishment, participants are less likely to use punishment when it costs the punisher 3 points rather than 1 point, and this was the case whether targets were relative cooperators or defectors. Based on these findings, our baseline hypothesis is that making punishment costly to punishers will reduce both prosocial punishment (punishing relative defectors) and antisocial punishment (punishing relative cooperators).

Hypothesis 1: Costly versus costless punishment under peer punishment. Peer punishers will engage less in prosocial and antisocial punishment when punishment is costly rather than costless.

Designated punishment may depart from such rational predictions. Compared to peer punishment, designated punishment heightens concerns about legitimacy, such as appearing proper, because punishment is centralized and delegated to one entity (e.g., leaders or outside authorities) who take on a unique role, commensurate with special status and power within each group, that elicits greater scrutiny (Devlin-Foltz and Lim 2008; Kosfeld et al. 2009). Under such conditions, and insofar as their power to punish is costly and viewed as proper, designated punishers will tend to use it in ways that are apparently proper and justifiable (prosocial) and avoid losing peer approval and support by abusing it because one defining effect of legitimacy is voluntary compliance with the demands of the situation (e.g., rules, norms, role expectations); people are more likely to comply with what they consider legitimate (Tyler 2006). In contrast, making punishment costless will reduce such normative constraints and let designated punishers gravitate toward antisocial punishment because it is no longer possible or meaningful to use such punishment to affirm or maintain propriety if the punishment itself lacks apparent propriety; it is unclear how to properly use something that is improper. In short, costly (vs. costless) punishment will prompt designated punishers to use punishment more prosocially.

Hypothesis 2: Costly versus costless punishment under designated punishment. Designated punishers will engage in more prosocial punishment (relative to antisocial punishment) when punishment is costly rather than costless.

We are not the first to examine the legitimacy of designated punishers (Baldassarri and Grossman 2011; Kosfeld et al. 2009). However, we are not aware of studies that examine the cost of punishment as the basis of legitimacy for designated punishers. We focused on the cost of punishment for both methodological and theoretical reasons. First, it is relatively easy to manipulate in both peer and designated punishment systems; alternative ways to legitimate enforcement, like electing punishers (e.g.,

Baldassarri and Grossman 2011), make little sense in peer punishment. Second, examining the cost of punishment is interesting because it highlights how the logic of legitimacy can diverge from the calculus of costly punishment. Although we are not the first to compare costly versus costless punishment or peer versus designated punishment, past research on the cost of punishment precluded costless punishment (Egas and Riedl 2008) or compared different forms of punishment (social vs. economic sanctions; Noussair and Tucker 2005), thus confounding the cost and form of punishment. To our best knowledge (see Balliet et al. 2011), no study has examined costly versus costless punishment by designated punishers.

METHODS

To test these ideas, our experiment modified the public goods game with punishment (Fehr and Gächter 2000) to manipulate the cost of punishment to enforcers (holding constant the cost to targets) in groups with peer versus designated punishment. In natural settings, punishment occurs in various forms, for example, formal versus informal, public versus private, material versus symbolic. It may be argued that punishment is never perfectly costless to the punisher, given administrative, psychic, relational, and various other nonmaterial costs (Adams and Mullen 2012). Our research is not designed to speak to this issue but to test the possibility that simply varying the material cost of punishment has consequences for how punishment is used.

To see whether perceptions of propriety varied across different enforcement systems, we examined propriety in three ways. First, participants completed an exit survey after the public goods games. The survey included questions about how proper people felt in their assigned roles during the public goods games. Second, we ran a pilot study in which volunteers read a written description of the public goods game with a different type of punishment (costly vs. costless and designated vs. peer) and rated the propriety of the enforcement system. Finally, we examined compliance, namely, the effects of punishment on cooperation, as a behavioral measure of legitimacy in the public goods experiment. By considering propriety at multiple levels, we view legitimacy as a property of an enforcement system as a whole rather than particular individuals, their acts, or their relations (Dornbusch and Scott 1975).

Participants and Procedure

Two hundred and thirty-four students (20.15 ± 1.90 years old, 36 percent male) from a large university were recruited for cash based on overall performance (average = \$12.50). The experiment was described as a study of organizational teamwork and took place in a laboratory with a no-deception policy. Participants were scheduled in groups of 6 to 12 but seated at isolated computer terminals. The entire experiment took place over the Internet through a custom website, thus preventing any face-to-face interactions or verbal communication and ensuring anonymity.

Prior to the experimental task, participants completed the consent form, detailed instructions, and a comprehension test. Next, they were randomly sorted into groups of three in different experimental conditions and assigned roles as punishers or non-punishers before completing "up to 10 rounds" of public goods games (the experiment ended after 6 periods). Finally, they were given an exit survey, received debriefing and payment, and dismissed.

Design and Materials

We created four experimental conditions: P1, P0, D1, and D0, where P and D designate peer versus designated punishment,

and 1 and 0 designate the cost of punishment. Our baseline condition (P1) was the standard public goods game with costly punishment (Fehr and Gächter 2000). In the first ("contribution") stage of each round, each member was given an endowment of 20 monetary units (MUs), of which they could contribute any amount to a team project or keep. Each MU contributed to the team project yielded a marginal per capita return of .5 MU for each member, and thus 1.5 MUs for the whole team, whereas keeping 1 MU yielded 1 MU for that member only. Thus, the earning π_{it} for member i in the first stage of period t is

$$\pi_{i,t}=20-c_{i,t}+.5\sum_{m=1}^{3}c_{m,t}. \qquad (1)$$

In the second ("punishment") stage, each punisher was given an opportunity to punish teammates, which entailed assigning "deduction" points (0 MUs to 10 MUs) out of one's own earnings to each other member. Each punishment point cost the punisher 1 MU and the target 3 MUs. Thus, the final payoff in each period t for player i in peer punishment is

$$\hat{\pi}_{i,t}=\pi_{i,t}-\sum_{i\neq m}^{3}p_{im,t}-3\sum_{i\neq m}^{3}p_{mi,t}. \qquad (2)$$

In the costless peer punishment condition (P0), the cost of punishment was 0 MU for punishers but 3 MUs for targets. In the two designated punishment conditions with costly (D1) or costless (D0) punishment, one member in each group was randomly chosen to be the sole punisher ("Leader") across all rounds. Punishers did not receive any additional endowment for punishment (O'Gorman et al. 2009).

To make the public goods game more engaging and meaningful to participants, it was described as a series of team projects in which members of an organization are asked to split their time between individual projects and team projects. Each "week," participants had 20 hours of unsupervised time (equivalent to 20 MUs) and earned different points from each hour contributed to team projects (depending on how much other members were contributing to team projects as well) versus individual projects, based on Equation 1. At the end of each week, punishers were given an opportunity to provide "feedback" to each other by assigning deduction points. The experimental materials are provided in Appendix A.

In our experiment, contributions and punishment were public knowledge. After each round, each member learned how much each person contributed, who was punished, by whom, and at what cost (in MUs) to the punisher and the punished teammate. This design was necessary to ensure comparability across peer and designated punishment conditions such that all punishers, not just designated punishers, were identifiable. It also served to reinforce the costliness of punishment, our key manipulation.

RESULTS

Table 1 presents summary statistics. Following the literature (Herrmann et al. 2008), we operationalized prosocial and antisocial punishment as punishing a relative defector versus cooperator, namely, a target who contributed fewer versus equal or more MUs than the punisher. Although punishers may engage in antisocial punishment for a number of different reasons, such as retaliation (Herrmann et al. 2008) or intergroup competition (Meier et al. 2012), we view antisocial punishment primarily as a display or abuse of power (Rand and Nowak 2011) since our experimental setup for the designated conditions rules out retaliation and intergroup competition.

We begin by comparing peer (Hypothesis 1) versus designated (Hypothesis 2)

Table 1. Means and Standard Deviations of Contributions, Punishment, and Earnings

	P0N = 61	P1N = 51	D0N = 21	D1N = 20
Average contribution per round (MUs)	10.92	10.85	7.63	9.11
	(3.35)	(3.18)	(2.74)	(2.70)
Total punishment (MUs)	14.74	5.16	7.81	8.85
	(20.00)	(10.07)	(11.51)	(7.43)
Total prosocial punishment (MUs)	8.08	3.14	3.57	7.70
	(13.16)	(4.45)	(4.64)	(6.84)
Total antisocial punishment (MUs)	5.17	2.02	4.24	1.15
	(8.50)	(7.08)	(9.34)	(1.95)
Total earning (MUs)	111.90	131.92	144.90	140.25
	(42.21)	(30.52)	(8.18)	(14.21)

Note: Standard deviations in parentheses. P0 = costless peer punishment, P1 = costly peer punishment, D0 = costless designated punishment, D1 = costly designated punishment, MUs = monetary units.

punishment separately because our hypotheses concern how punishment cost affects the use of punishment under different punishment regimes. We then introduce econometric tests that compare peer and designated punishment more directly. After the hypothesis tests, we present evidence that varying the cost of punishment changed the perceived propriety of punishment. All tests are two-tailed.

Effects of Propriety on Punishment

As expected, peer punishers punished less when punishment was costly rather than costless, but this was not the case for designated punishers. Over six periods, peer punishers used less prosocial punishment in P1 (M = 3.13 MUs, SD = 4.45) than P0 (M = 8.08 MUs, SD = 13.16), $t(109) = 2.56$, $p = .01$, and antisocial punishment in P1 (M = 2.02 MUs, SD = 7.08) than P0 (M = 5.17 MUs, SD = 8.50), $t(109) = 2.10$, $p = .04$. In contrast, costly punishment increased prosocial punishment by designated punishers from 3.57 MUs (SD = 4.64) to 7.70 (SD = 6.83), $t(39) = 2.27$, $p = .03$, whereas it decreased antisocial punishment from 4.24 MUs (SD = 9.34) to 1.15 (SD = 1.95), $t(39) = 1.45$, $p = .16$, although this effect did not reach significance at the 5 percent level.

While these patterns are consistent with our hypotheses, the descriptive results may be biased because they do not control for the effects of contributions on punishment or the repeated measures nested in individuals and teams. It is also difficult to compare peer versus designated punishment directly because of methodological differences (e.g., 1 vs. 3 punishers). We addressed these issues as follows by using errors clustered at the level of individual punishers and teams and controlling for contributions from the punisher, the target, and the team total in each round and the fixed effects of rounds:

$$p_{im,t} = B_0 + B_1 \sum_{m=1}^{3} c_{m,t} + B_2 c_{i,t} + B_3 (c_{i,t} - c_{m,t}) + B_4 \sum_{m=1}^{3} p_{im,t-1} + B_{5\ldots} Treatment. \quad (3)$$

The subscript m indicates punisher 1 . . . 3 (only 1 under designated punisher), i denotes rounds 1 . . . 6, and t is the current round. Thus, p_{imt} is punishment by i to m, and c_{it} is contribution by i, both in round t. In this model, B_0 is the constant, B_1 is the team's total contribution, B_2 is the punisher's contribution, B_3 is the difference in contribution between the punisher and a target, B_4 is punishment given in $t - 1$,

and B_5 . . . are the types of punishment (costly vs. costless, peer vs. designated, and their interaction effect). Following the literature (e.g., Ashley, Ball, and Eckel 2010), we submitted this model to tobit regression because the dependent variable, punishment per target and round in MUs, is censored at 0 and 10 MUs.

The results (Table 2) converge with the descriptive patterns. In peer punishment groups, costly punishment shows negative effects on both prosocial punishment, $B = -2.39$, robust SE = .67, $p < .001$, and antisocial punishment, $B = -3.46$, robust SE = 1.60, $p = .03$, supporting Hypothesis 1. Under designated punishment, in comparison, costly punishment increased prosocial punishment, $B = 1.14$, robust SE = .57, $p = .047$, while it had no effect on antisocial punishment, $B = .67$, robust SE = 1.50, $p = .65$, consistent with Hypothesis 2. Finally, in the pooled data, the effect of costly $\times$ designated punishment on prosocial punishment was significant and positive, $B = 3.40$, robust SE = 1.01, $p = .001$. Looking at antisocial punishment, we find a negative main effect of costly punishment, $B = -3.50$, robust SE = 1.61, $p = .03$, but no interaction effect, $B = 2.80$, robust SE = 2.53, $p = .27$. Altogether, these results show that costly punishment increased prosocial punishment by designated punishers but not peer punishers and reduced antisocial punishment under both peer and antisocial punishment.[1]

Evidence of Propriety

Overall, our results support our reasoning that imbuing punishment with propriety helps regulate the use of power by motivating prosocial punishment without increasing antisocial punishment by designated punishers. An alternative explanation, however, is that the cost of punishment changed its credibility, not propriety. That is, punishment may feel more credible—the punisher really means it—if it is costly because, according to signaling theory (Spence 1974), signals are taken more seriously if they are costly. Creditability and propriety are different because both proper and improper acts can be credible (e.g., a mobster threatening a person's life). To address this issue, we provide three lines of evidence for propriety.

First, prior to the laboratory experiment, we ran a pilot study in which 240 volunteers from Amazon Mechanical Turk (all North Americans; 3 did not finish the study) were recruited in exchange for monetary compensation and asked to provide feedback on a new experiment "designed to examine teamwork." Participants were randomly assigned to read a description of a public goods game with a different type of punishment (costly vs. costless and peer vs. designated), taken from the main experiment (see Appendix B). Next, participants answered two questions about propriety: "How fair/appropriate is this enforcement system?" (1 = very unfair/inappropriate, 5 = very fair/appropriate; Spearman's $\rho > .79$).[2] Designated punishment was perceived as more fair and appropriate when costly (M = 4.16, SD = 1.26) than costless (M = 3.48, SD = 1.34), $t(120) = 2.84$, $p = .005$. However, peer punishment was

[1]As robustness checks, we examined punishment frequency (whether punishment occurred or not) and severity (points assigned, given actual punishment). For frequency, we obtained similar results as Table 2. For severity, we did not obtain robust results since the analysis considers instances of actual punishment only, reducing the sample sizes. These results are available on request.

[2]It is worth noting that propriety is more than fairness because, despite the high correlation here, fairness and appropriateness are conceptually distinct. For instance, something fair can be inappropriate (e.g., equal division of illicit money). This is crucial because fairness alone cannot explain why designated punishers engaged in more prosocial punishment since it is possible to be "fair" in other ways, for example, by not engaging in punishment at all.

Table 2. Predictors of Punishment Per Punisher Per Round

	Peer punishment		Designated punishment		Pooled data	
	Prosocial	Antisocial	Prosocial	Antisocial	Prosocial	Antisocial
Costly punishment	–2.39**	–3.46*	1.14*	.67	–2.32**	–3.50*
	(.67)	(1.60)	(.57)	(1.50)	(.61)	(1.61)
Designated punishment					–1.47*	–1.95
					(.73)	(1.73)
Costly × designated					3.40**	2.80
					(1.01)	(2.53)
Punisher contribution	–.32**	–.41*	–.46**	–.46	–.36**	–.31
	(.12)	(.19)	(.16)	(.31)	(.10)	(.17)
Punisher – target contribution	.75**	–.11	.82**	.66**	.76**	.05
	(.09)	(.11)	(.11)	(.22)	(.07)	(.10)
Group mean contribution	.38*	.55*	.23	–.22	.29*	.22
	(.15)	(.24)	(.18)	(.35)	(.12)	(.21)
Constant	–1.62	–8.41**	–.15	2.43	–.38	–4.67**
	(1.11)	(2.02)	(.88)	(2.03)	(.88)	(1.75)
Team level intercept	2.17	12.18*			1.77*	13.28*
	(1.12)	(6.03)			(.80)	(5.35)
Individual level intercept	7.23**	12.86**	3.39**	6.15**	6.18**	12.97**
	(.81)	(1.92)	(.58)	(1.74)	(.58)	(1.72)
N	517	817	190	302	707	1119
Log likelihood	–696.82	–522.30	–261.48	–150.58	–974.35	–696.06

Note: Results from tobit regression with errors clustered at the individual and team levels. Robust standard errors in parentheses.
*$p < .05$. **$p < .01$, two-tailed tests.

perceived as equally fair and appropriate when costly (M = 4.23, SD = 1.06) versus costless (M = 4.34, SD = 1.37), $t(115) = .49$, $p = .62$. Thus, the cost of punishment changed perceptions of propriety at the system level but only under designated punishment.

Second, in the exit survey after the public goods games, participants were asked how proper (1 = very unfair/disrespected, 7 = very fair/respected, Spearman's ρ = .48) they felt while playing their assigned roles.[3] Designated punishers reported feeling more legitimate in D1 (M = 4.88, SD = 1.73) than in D0 (M = 3.55, SD = 1.50), $t(39) = 2.63$, $p = .01$. Non-punishers also felt more proper in D1 (M = 4.61, SD = 1.21) than D0 (M = 3.57, SD = 1.48), $t(80) = 3.47$, $p = .008$. These results were robust to controlling for total punishment received and final earning. No such patterns were found between P0 (M = 3.78, SD = 1.42) and P1 (M = 4.14, SD = 1.52), $t(109) = 1.30$, $p = .20$. In addition, in the designated punishment conditions only, non-punishers were asked how fair the punisher in their team was. They evaluated their punisher as more fair in D1 (M = 4.53, SD = 1.95) than D0 (M = 3.43, SD = 1.95), $t(80) = 2.54$, $p = .01$.

Third, an indirect but consequential measure of legitimacy is compliance, namely, how much members increase

[3]Rather than simply replicating the vignette study, we changed the level of analysis from the system as a whole to individual roles within the system to see if manipulating legitimacy at the system level would affect individual behaviors and experiences at the role level. In doing so, we used the term *respected* rather than *appropriate* to better capture how people think they are seen by others.

Table 3. Effects of Punishment on Compliance

	Designated	Peer
Own contribution, $t - 1$	–.73**	–.67**
	(.09)	(.09)
Group average contribution, $t - 1$	.58**	.70**
	(.12)	(.08)
Punisher contribution, $t - 1$	.07	
	(.06)	
Punishment received, $t - 1$	.03	.03
	(.12)	(.07)
Costly punishment	–.00	–.36
	(.28)	(.23)
Punishment received × costly	.36*	.15
	(.16)	(.14)
Constant	.21	.74
	(.59)	(.50)
N	410	570
Log likelihood	–955.25	–1,321.08

Note: Results from tobit regression with errors clustered at the individual level. Dependent variable is contribution in monetary units (MUs) in round *t*. Robust standard errors in parentheses.
*$p < .05$. **$p < .01$, two-tailed tests.

their contributions after receiving punishment (Baldassarri and Grossman 2011; Zelditch 2001). If costly punishment is more credible but not legitimate, we should find costly punishment to increase compliance after receiving either prosocial or antisocial punishment. To the contrary, we find that costly punishment increased the efficacy of prosocial punishment only and only for designated punishers.

Table 3 shows results from tobit regression predicting contributions in MUs in round t as a function of punishment received in $t - 1$. We found a significant positive effect of punishment received in $t - 1 \times$ costly punishment under designated punishment, $B = .36$, robust SE = .16, $p = .03$, but not peer punishment, $B = .15$, robust SE = .14, $p = .28$, indicating that contributions increased more after receiving costly (vs. costless) punishment from designated punishers but not from peer punishers.

An alternative explanation for the difference in compliance under designated punishment is that punishers engaged in more antisocial punishment under costless punishment, which perhaps alienated group members and reduced their compliance. In other words, compliance dropped not because punishment was costless (and therefore less proper) but because punishers were abusing it. To consider this issue, we re-specified our econometric model by replacing the term *punishment received* in Table 3 with *prosocial punishment received* and *antisocial punishment received* and ran the new model for the costless and costly designated punishment conditions separately. This model should help reveal differences in compliance under costless versus costly punishment due specifically to antisocial punishment.

Table 4 shows the results. First, contrary to the idea that punishers drove down compliance by engaging in antisocial punishment, antisocial punishment had no effect on compliance in either condition. Second, prosocial punishment is positive and marginally significant under costly punishment, $B = .93$, SE = .54,

Table 4. Effects of Punishment on Compliance under Designated Punishment

	Costless	Costly
Own contribution, $t-1$	–.66**	–.87**
	(.12)	(.13)
Group average contribution, $t-1$	.36+	.70**
	(.18)	(.18)
Punisher contribution, $t-1$	.32**	–.05
	(.11)	(.07)
Prosocial punishment received, $t-1$	–.23	.93+
	(.51)	(.54)
Antisocial punishment received, $t-1$	.77	–.01
	(.75)	(.67)
Constant	–.44	1.69
	(.63)	(1.17)
N	210	200
Log likelihood	–436.17	–497.24

Note: Results from tobit regression with errors clustered at the level of individual targets of punishment. Robust standard errors in parentheses.
$+p < .1$. $**p < .01$, two-tailed tests.

$p = .08$, but not under costless punishment, $B = –.23$, SE = .51, $p = .65$. Thus, prosocial punishment lost its efficacy when it was made costless. While post hoc, these results provide support for the idea that making punishment costless affected compliance directly (by reducing propriety) rather than indirectly (by increasing antisocial punishment).[4]

Cooperation and Efficiency

As supplemental analysis, we examined cooperation and efficiency across conditions (Figure 1). Censored regression of individual contribution per round on experimental conditions with random effects at the individual participant level and team level found a positive effect of costly punishment on average contribution under designated punishment, $B = 1.48$, robust SE = .66, $p = .026$. We found no difference between P0 and P1, $B = .59$, robust SE = 1.03, $p = .57$, although peer punishment produced more cooperation than designated punishment under both costly punishment, $B = 1.59$, robust SE = 79, $p = .044$, and costless punishment, $B = 3.62$, robust SE = .88, $p < .001$.

Next, we regressed individual earning in MUs per round on conditions, with random effects at the individual and team levels. The results reveal that earnings were higher under costly punishment, $B = 3.02$, SE = 1.40, $p = .031$, and under designated punishment, $B = 3.44$, SE = 1.93, $p = .009$, but there was no effect of costly $\times$ designated punishment, $B = -2.74$, robust SE = 1.93, $p = .16$. Figure 1b suggests that these patterns are driven by the low earnings in groups with costless peer punishment, which showed the highest levels of punishment. Prior research has found that costly peer punishment increases cooperation, but overall efficiency gains are erased by punishment (Egas and Riedl 2008; Herrmann et al. 2008), at least in experiments with

[4]It is curious that punisher contribution in $t-1$ is significant under costless punishment only. Our interpretation is that when punishment was costless and deemed illegitimate, group members paid greater attention to punisher contributions.

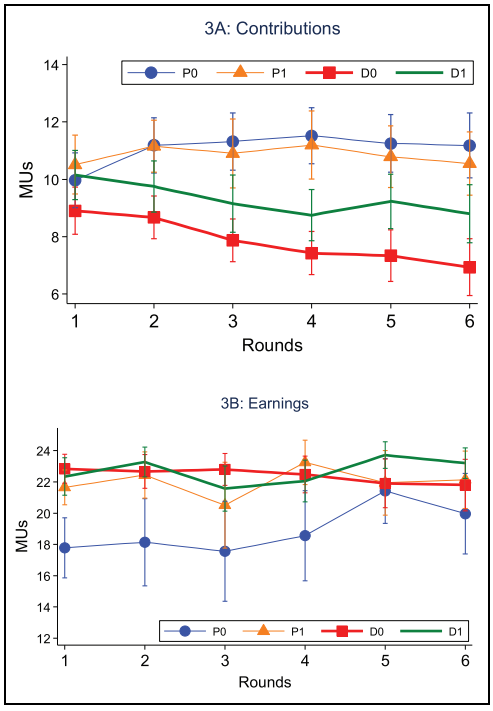


Figure 1. Individual contributions and earnings by condition.
Note: P0 = costless peer punishment, P1 = costly peer punishment, D0 = costless designated punishment, D1 = costly designated punishment.

short time horizons (6–8 rounds; Gächter et al. 2008). Our research suggests that this may be the case in particular when peer punishment is costless. In contrast, costly designated punishment increased cooperation without reducing efficiency, relative to costless designated punishment.[5]

DISCUSSION

An enduring insight from sociology is that legitimacy exerts powerful constraints on actors to comply with prevailing norms of conduct. For instance, Zelditch and Walker (2000) suggest that concerns about legitimacy can help regulate the use of power by invoking a sense of what is morally right or socially proper for power-holders (also Fast and Chen 2009; Kuwabara et al. 2016). The present research extends this argument on the basis of legitimacy induced by making the punishment costly rather than costless to the punisher. Consistent with our hypotheses, designated punishers engaged in significantly more prosocial punishment but not antisocial punishment when punishment was costly and thus perceived as proper. In contrast, peer punishers were less likely to use costly rather than costless punishment. We also found that costly punishment was more effective than costless punishment in sustaining cooperation, but only under designated punishment.

Our results for designated punishers are related to the effect of explicit cost to invoke norms of economic rather than social exchange. Shampanier, Mazar, and Ariely (2007) found that when offered candies at 1 cent each, students took four on average; when offered free candies at 0 cent each, more students took candies, but almost none took more than one. The argument is that even a small cost can invoke a market mindset that helps justify and motivate consumption, whereas zero cost invokes social norms against taking more than one's share. Enforcement may be subject to a similar psychology (Tenbrunsel and Messick 1999), invoking different norms when it is costly rather than costless. In our experiment, simply changing the cost of punishment from 0 MU to 1 MU amounted to a noticeable shift from costless to costly punishment that was not prohibitively large in economic terms yet salient enough to alter its moral significance, imbuing acts of punishment with legitimacy and reducing antisocial punishment.

[5]O'Gorman, Henrich, and Van Vugt (2009) found that costly designated punishment sustains as much cooperation as costly peer punishment. One key difference in our designs is that participants stayed in the same group across all rounds instead of rotating after every round. Peer punishment (but not designated punishment) may be more effective in fixed groups.

These patterns call into question what the cost of punishment really represents in experiments on costly punishment. Before the recent surge of interest in costly punishment, exchange theorists pursued a productive line of work on the use of coercive power without explicit cost (Lawler, Ford, and Blegen 1988; Molm 1997). The experiment by Fehr and Gächter (2000) made a departure by assuming costly punishment to provide a more stringent test of the idea that people are willing to use punishment even at their own cost. As our research suggests, however, punishment cost is more than a price; it is also a social signal that changes how people interpret the act of punishment. Indeed, many acts of punishment in real life do not come with an explicit price tag. An important direction for future research is to better understand conditions under which punishment is actually viewed as costly or costless.

Our findings have implications beyond the laboratory. For instance, recent years have seen a phenomenal growth of reputation systems designed to regulate online markets (e.g., eBay, Amazon, Yelp) by harnessing peer-to-peer enforcement that is virtually costless. Despite the success of these systems, however, a persistent challenge is how to ensure prosocial enforcement, namely, feedback that is viewed as proper and legitimate (Resnick et al. 2000). Our research suggests that making feedback a little more costly may help curtail antisocial punishment by raising not only the economic cost of punishment but also concerns about legitimacy.

Although legitimacy may derive from various sources, the idea that legitimacy may also inhere in the cost of punishment is empirically novel and may shed light on the problem of enforcement. If costly punishment is a source of legitimacy in itself, people might engage in enforcement precisely because it is a costly and therefore effective signal of their prosociality and social status (Barclay 2006; Jordan et al. 2016). This points to conditions under which punitive sentiments may have evolved: in hierarchical groups that recognize selfless punishers as legitimate leaders (Traulsen, Röhl, and Milinski 2012). Such groups have received relatively scant attention in the literature on costly punishment even though flat groups with no clear hierarchical differentiation are rare outside of the laboratory (Gruenfeld and Tiedens 2010). Hierarchies emerge quickly and spontaneously, creating differentiations in power and status that become the basis of social arrangements like designated punishment. Our understanding of how groups enforce cooperation in hierarchical groups is incomplete without greater efforts to account for the social psychology of power and legitimacy. In particular, more work is needed to better understand the conditions under which costly versus costless punishment evolved in different types of groups.

APPENDIX

APPENDIX A: INSTRUCTIONS FOR THE PUBLIC GOODS EXPERIMENT

Welcome Page

Welcome to the Behavioral Research Lab. Today, you and other participants will be participating in a "Virtual Teams" study. You are one of up to 15 participants who will be participating in this session. Momentarily, you will be interacting with some of them over the Internet in a series of simple group tasks that determines your cash earning from this experiment. The entire study lasts less than 45 minutes and will be completed from your computer.

Next, we ask you to carefully read the instructions that explain the experiment. For the remainder of the session, please do not use the computer for any purpose besides the study, and refrain from hitting the "Back" button, as this may disrupt the flow of the experiment. Please also turn off your phone, TV, radio, or any other device that may distract you.

Momentarily, your Research Assistant will tell you and other participants to proceed at the same time. Until then, please do NOT proceed. Thank you for your patience.

Page 1: Instruction

About Today's Study: Modern organizations consist of employees who work alone as well as in teams to achieve a wide range of personal and collective goals. This study is designed to examine how people work in teams to complete a series of "virtual team projects." Imagine that you are summer interns at the organization. After reading the instructions, you and other participants will be sorted into different teams, each consisting of three members. Throughout the entire experiment, you will stay in your assigned team. Teams will work on their projects independently and work at their own pace.

Page 2: Instruction

Team vs. Personal Projects: In this organization, you and other interns have 20 hours every week that you can decide to spend on personal projects or team projects. Your task is to decide on your own how many hours to work on, or "contribute" to, team projects vs. personal projects each week. You will earn bonus points (BP) from both team and personal projects, depending on how you and other teammates allocate their time to personal vs. team projects. Your individual goal is to earn as many points from working on both team and personal projects.

Every hour you spend on the team project benefits the entire team, but each hour you spend on your own projects benefits you even more. Specifically, each team member's earning from a team project is: (20 hours – Hours you spend on team projects) + 50 percent of (Total hours you and your teammates spend on team projects). For instance, if you spend 12 hours, and your teammates spend 10 hours each, your earning is $20 - 12 + .5 \times (10 + 10 + 12) = 24$ BP.

Don't worry, we will handle all the math for you during the game and show you the results each week. The **important thing to remember** is that 1) your earning depends on what you and your teammates decide to work on each week, and 2) spending time on personal projects is 50% more profitable to you personally than contributing to team projects. However, working on team projects is the most profitable if everyone contributes maximally.

Page 3: Instruction

Contributing to Your Team: In your team, you will be asked to complete up to 10 rounds or "weeks" of work. In each week, you will have 60 seconds to decide how much to work on team vs. personal projects. Because your teammates are waiting for their turn to work on their projects, it is vital that you make your decisions in a timely manner during the experiment. If you miss your turn, the server will automatically make a contribution to team projects based on your previous contribution so your teammates are not affected by your mistake, but your personal earning will be 0 BP for the week.

After each team project, we will calculate and show you the results of your decisions: how many hours you and your teammates contributed to the project, and what each of you earned. For example:

Member	Worked for team	Total earning	Assign deduction pts
Player A1	14 hrs	23.5 BP = (20-14) + .5*(14+9+12)	0 DP
Player B2	9 hrs	28.5 BP = (20-9) + .5*(14+9+12)	0 DP
Player C3 (you)	12 hrs	25.5 BP = (20-12) + .5*(14+9+12)	

Feedback: Although you are not allowed to talk to your teammates during the experiment, giving feedback is an important aspect of team work in organizations. Each week, one [each] of you in your team will be allowed to assign between 0 to 10 **deduction points (DP)** to each teammate. Each deduction point (1 DP) will cost you 1 [0] BP out of your week's earning, and it will cost the teammate 3 BP out of his or her week's earning. Everyone in the team will know that assigning deduction points is therefore not entirely free [totally costless]. In this organization, giving deduction points is an important way of letting others know how you feel about their contributions.

Page 4: Instruction

Your Responsibility: The entire study takes place online, through your computer. In compliance with the Internal Review Board, there is no deception in this study; that means you will be interacting with actual human participants, not computer robots. Therefore, it is important that you complete the entire session so that you do not prevent others from completing the study. If one person drops out, other participants may be forced to quit the session as well. If you feel like your browser is stuck during the experiment, please let the RA know before hitting the Refresh or Back buttons.

Your Privacy: We take your privacy seriously. We promise anonymity and confidentiality of all information you provide. During the experiment, you will be identified by a numerical ID only, and we will not release your personal information in any identifiable way.

Page 5: Instruction

Your Compensation: For giving us your valuable time and full attention during today's study, your compensation is a cash payment based on your final earnings at the end of the study. Each point you earn is worth 10 cents. If your earning is negative at the end of the experiment, you will earn $7 for completing the entire study.

We are almost ready to begin the study. On the next page, we ask you to complete short questionnaires, which will test your understanding of the instructions.

Page 6: Control Questions

Before we begin: We need to make sure everyone understands the experiment. Please answer the following questions. Once everyone completes this quiz, we will assign you to teams and begin the experiment. [This page did not proceed until all questions were answered correctly.]

- If your two teammates contribute a total of 24 hours, and you contributed 10 hours, what is your total earning from contributions for the week?
- If Player A contributed 16 hours, B contributed 5 hours, and C contributed 10 hours to team projects, who earned the most?
- True or false: Each deduction point will cost the giver 1 point and the receiver 3 BP.
- How many "weeks" of projects will each team complete?

Role Assignment

Waiting for everyone to finish the Pop Quiz. Thank you for your patience.

[Once everyone finishes the quiz] The server has assigned you (Player 14D) to Team 56 with Player 7B and 11R.

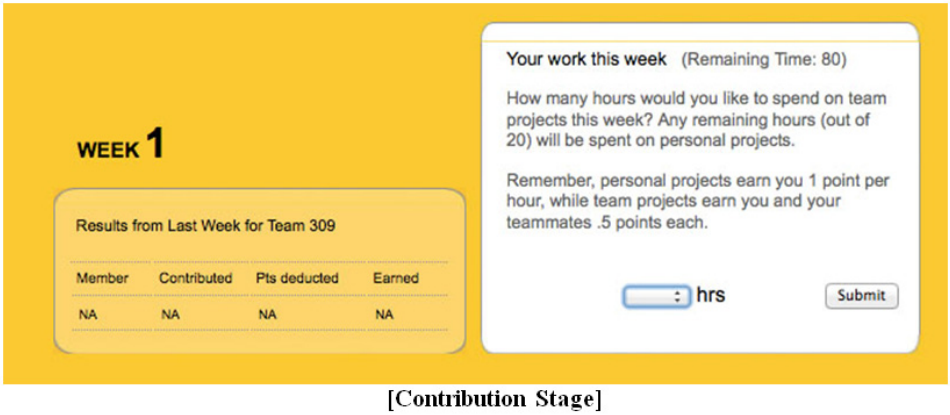
[Peer punishment] In your team, everyone is able to provide feedback by assigning deduction points to each other after each round.

[Designated punishment] Based on the day of your birthdays, the server has **randomly** selected you to be the Leader. This means you are the only person who was chosen by chance to assign deduction points (DP) after each round.

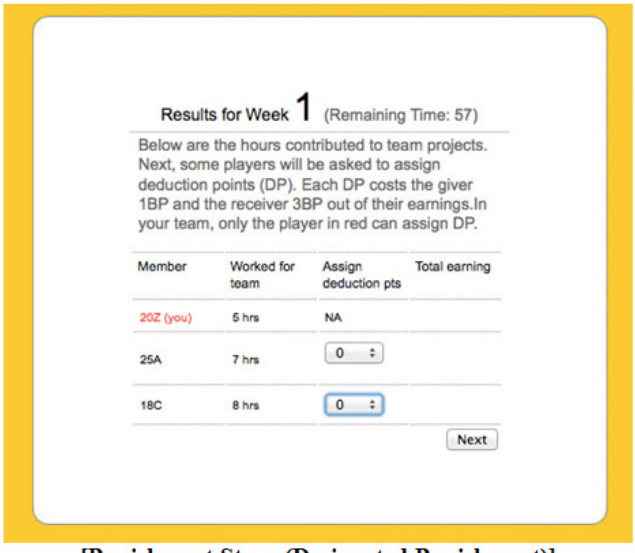
When you are ready, please proceed to start Week 1.

Screenshots of Rounds

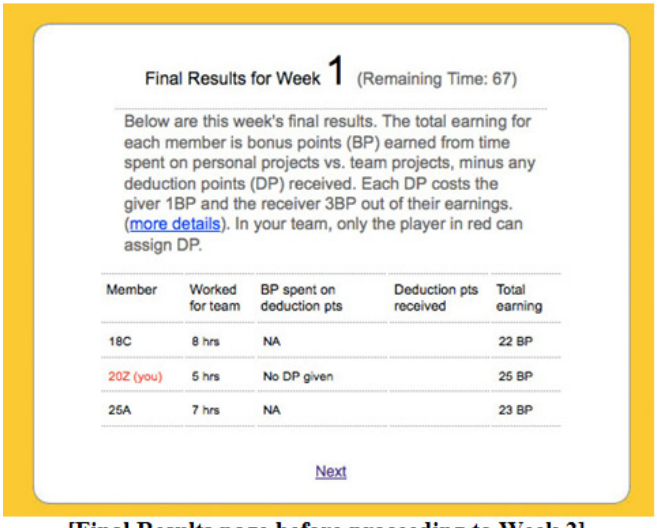
[Each round ("week") consisted of 3 pages: contribution, punishment, and the final results.]



[Contribution Stage]



[Punishment Stage (Designated Punishment)]



[Final Results page before proceeding to Week 2]

Exit Survey [After 6 Rounds]

Congratulations. You and your team have completed the experiment. Your personal earning is ___ points ($___).

In order to protect the anonymity of all participants, everyone will be dismissed from the lab at the same time after everyone has finished. This will also give the RA time to prepare your payment.

In the meanwhile, we want to know how you feel about your experience in the study. Please reflect back on the study carefully and respond to the following questions.

- Overall, how satisfied are you with your own performance? [1 = very unsatisfied, 7 = very satisfied]
- Overall, how satisfied are you with your team performance? [1 = very unsatisfied, 7 = very satisfied]
- How engaged or unengaged did you feel during the experiment? [1 = very unengaged, 7 = very engaged]

- How fair or unfair did you feel about performing your role? [1 = very unfair, 7 = very fair]
- How respected or disrespected did you feel about performing your role? [1 = very disrespected, 7 = very respected]

APPENDIX B: INSTRUCTIONS FOR THE PILOT STUDY

Page 1

We are a team of academics designing a psychology experiment on teamwork, and we are interested in getting feedback on our design before it is finalized. Please carefully read the description of the design below and answer a few questions for us. Thank you very much for your cooperation.

Page 2

Modern organizations consist of employees who work alone as well as in teams to achieve a wide range of personal and collective goals. This experiment is designed to examine how people work in teams to complete a series of "team projects." Participants will be asked to imagine that they are summer interns in an organization. Before they start, they will be sorted into different teams, each consisting of three members.

In this experiment, participants will be asked to complete several rounds of work. Each round represents a week of internship, where interns will have 20 hours that they can decide to spend on personal projects or team projects. Their task is to decide on their own how many hours to "contribute" each week to team projects or personal projects. Participants will earn Bonus Points from both team and personal projects, depending on how they and other teammates allocate their time to personal vs. team projects.

Specifically, each team member's earning from a team project is: (20 hours – Hours he or she spends on team projects) + 50% of (Total hours he or she and his or her teammates spend on team projects). For instance, if he or she spends 12 hours, and his or her teammates spend 10 hours each, her earning is $20 - 12 + .5 \times (10 + 10 + 12) = 24$ Bonus Points. Thus, it is always more profitable for each person to spend time on personal projects, but working on team projects is the most profitable as long as everyone contributes maximally.

Page 3

[Costly designated punishment condition]

To ensure team work, we will randomly choose one team member to be the "enforcer" who can assign between 0 to 10 Deduction Points to each teammate. After each work week, the enforcer will see how many hours each intern spent on personal vs. team projects and assign Deduction Points as he/she wishes. Each deduction point (1 DP) will reduce 3 Bonus Points out of an intern's earning, but it will also cost the enforcer 1 Bonus Point out of his/her own earning. Everyone in the team will know that assigning deduction points is therefore not entirely free to the enforcer. The enforcer is the only person who can assign deduction points.

Now, imagine that you were randomly chosen to be an enforcer to assign deduction points to your teammates at your own expense.

[Costless designated punishment condition]

To ensure team work, we will randomly choose one team member to be the "enforcer" who can assign between 0 to 10 Deduction Points to each teammate. After each work week, the enforcer will see how many hours each intern spent on personal vs. team projects and assign Deduction Points as he/she wishes. Each deduction point (1 DP) will reduce 3 Bonus Points out of an intern's earning, but it will cost the enforcer nothing at all (0 Bonus Point). Everyone in the team will know that assigning deduction points is totally costless to the enforcer. The enforcer is the only person who can assign deduction points.

Now, imagine that you were randomly chosen to be an enforcer to assign deduction points to your teammates at no expense to yourself.

[Costly peer punishment]

To ensure team work, each team member will be an "enforcer" who is able to assign between 0 to 10 Deduction Points to each other. After each work week, each team member will see how many hours each person spent on personal vs. team projects and will be asked to assign Deduction Points to each person as he/she wishes. Assigning each deduction point (1 DP) will cost 1 Bonus Point, but will deduct 3 Bonus Points out of the target's earning. Everyone in the team will know that assigning deduction points is NOT entirely free costless.

Now, imagine that you are one of the members in this study.

[Costless peer punishment]

To ensure team work, each team member will be an "enforcer" who is able to assign between 0 to 10 Deduction Points to each other. After each work week, each team member will see how

many hours each person spent on personal vs. team projects and will be asked to assign Deduction Points to each person as he/she wishes. Assigning each deduction point (1 DP) will cost nothing at all, but will deduct 3 points out of the target's earning. Everyone in the team will know that assigning deduction points is totally costless.

Now, imagine that you are one of the members in this study.

Page 4

Based on what you just read, please answer the following questions.

1. How fair is this enforcement system? [5 = Very fair,1 = Very unfair]
2. How appropriate is this enforcement system? [5 = Very appropriate, 5 = Very inappropriate]

REFERENCES

Adams, Gabrielle S., and Elizabeth Mullen. 2012. "The Social and Psychological Costs of Punishing." *Behavioral and Brain Sciences* 35:15–16.

Ashley, Richard, Sheryl Ball, and Catherine Eckel. 2010. "Motives for Giving: A Reanalysis of Two Classic Public Goods Experiments." *Southern Economic Journal* 77:15–26.

Baldassarri, Delia, and Guy Grossman. 2011. "Centralized Sanctioning and Legitimate Authority Promote Cooperation in Humans." *Proceedings of the National Academy of Sciences* 108:11023–27.

Balliet, Daniel, Laetitia B. Mulder, and Paul A. M. Van Lange. 2011. "Reward, Punishment, and Cooperation: A Meta-Analysis." *Psychological Bulletin* 137:594–615.

Barclay, Pat. 2006. "Reputational Benefits for Altruistic Punishment." *Evolution and Human Behavior* 27:325–44.

Biggart, Nicole Woolsey, and Gary G. Hamilton. 1984. "The Power of Obedience." *Administrative Science Quarterly* 29: 540–49.

Blau, Peter M. 1964. *Exchange and Power in Social Life*. New Brunswick, NJ: Transaction.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton, NJ: Princeton University Press.

Chaudhuri, Ananish. 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature." *Experimental Economics* 14:47–83.

Devlin-Foltz, Zack, and Katherine Lim. 2008. "Responsibility to Punish: Discouraging Free-Riders in Public Goods Games." *Atlantic Economic Journal* 36:505–18.

Dornbusch, Sanford M., and W. Richard Scott. 1975. *Evaluation and the Exercise of Authority*. San Francisco: Jossey-Bass Publishers.

Egas, Martijn, and Arno Riedl. 2008. "The Economics of Altruistic Punishment and the Maintenance of Cooperation." *Proceedings of the Royal Society B: Biological Sciences* 275:871–78.

Fast, Nathanael J., and Serena Chen. 2009. "When the Boss Feels Inadequate: Power, Incompetence, and Aggression." *Psychological Science* 20:1406–13.

Fast, Nathanael J., Nir Halevy, and Adam D. Galinsky. 2012. "The Destructive Nature of Power without Status." *Journal of Experimental Social Psychology* 48:391–94.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90:980–94.

Fehr, Ernst, and Bettina Rockenbach. 2003. "Detrimental Effects of Sanctions on Human Altruism." *Nature* 422:137–40.

Gächter, Simon, Elke Renner, and Martin Sefton. 2008. "The Long-Run Benefits of Punishment." *Science* 322:1510.

Gruenfeld, Deborah H., and Larissa Z. Tiedens. 2010. "Organizational Preferences and Their Consequences." Pp. 1252–87 in *Handbook of Social Psychology*, edited by S. T. Fiske, D. T. Gilbert, and G. Lindzey. New York: John Wiley & Sons, Inc.

Heckathorn, Douglas D. 1990. "Collective Sanctions and Comliance Norms: A Formal Theory of Group-Mediated Social Control." *American Sociological Review* 55:366–84.

Hegtvedt, Karen A., and Cathryn Johnson. 2009. "Power and Justice: Toward an Understanding of Legitimacy." *American Behavioral Scientist* 53:376–99.

Herrmann, Benedikt, Christian Thoni, and Simon Gächter. 2008. "Antisocial Punishment across Societies." *Science* 319: 1362–67.

Horne, Christine, and Anna Cutlip. 2002. "Sanctioning Costs and Norm Enforcement: An Experimental Test." *Rationality and Society* 14:285–308.

Johnson, Cathryn, Timothy J. Dowd, and Cecilia L. Ridgeway. 2006. "Legitimacy as a Social Process." *Annual Review of Sociology* 32:53–78.

Jordan, Jillian J., Moshe Hoffman, Paul Bloom, and David G. Rand. 2016. "Third-Party Punishment as a Costly Signal of Trustworthiness." *Nature* 530:473–76.

Kiyonari, Toko, and Pat Barclay. 2008. "Cooperation in Social Dilemmas: Free Riding May Be Thwarted by Second-Order Reward Ratehr Than by Punishment." *Journal of Personality and Social Psychology* 95:826–42.

Kosfeld, Michael, Akira Okada, and Arno Riedl. 2009. "Institution Formation in Public Goods Games." *The American Economic Review* 99:1335–55.

Kuran, Timur. 1995. "The Inevitability of Future Revolutionary Surprises." *American Journal of Sociology* 100:1528–51.

Kurzban, Robert, Peter DeScioli, and Erin O'Brien. 2007. "Audience Effects on Moralistic Punishment." *Evolution and Human Behavior* 28:75–84.

Kuwabara, Ko, Siyu Yu, Alice J. Lee, and Adam D. Galinsky. 2016. "Status Decreases Dominance in the West but Increases Dominance in the East." *Psychological Science* 27:127–37.

Lawler, Edward J., Rebecca S. Ford, and Mary A. Blegen. 1988. "Coercive Capability in Conflict: A Test of Bilateral Deterrence Versus Conflict Spiral Theory." *Social Psychology Quarterly* 51:93–107.

Meier, Stephan, Lorenz F. Goette, David Huffman, and Matthias Sutter. 2012. "Group Membership, Competition, and Altruistic Versus Antisocial Punishment: Evidence from Randomly Assigned Army Groups." *Management Science* 58:948–60.

Molenmaker, Welmer E., Erik W. de Kwaadsteniet, and Eric van Dijk. 2016. "The Impact of Personal Responsibility on the (Un) Willingness to Punish Non-Cooperation and Reward Cooperation." *Organizational Behavior and Human Decision Processes* 134:1–15.

Molm, Linda D. 1997. *Coercive Power in Social Exchange*. Cambridge, UK: Cambridge University Press.

Noussair, Charles, and Steven Tucker. 2005. "Combining Monetary and Social Sanctions to Promote Cooperation." *Economic Inquiry* 43:649–60.

O'Gorman, Rick, Joseph Henrich, and Mark Van Vugt. 2009. "Constraining Free Riding in Public Goods Games: Designated Solitary Punishers Can Sustain Human Cooperation." *Proceedings of the Royal Society B: Biological Sciences* 276:323–29.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge UK: Cambridge University Press.

Rand, David G., and Martin A. Nowak. 2011. "The Evolution of Antisocial Punishment in Optional Public Goods Games." *Nature Communications* 2:434.

Resnick, Paul, Richard Zeckhauser, Eric Friedman, and Ko Kuwabara. 2000. "Reputation Systems." *Communications of the ACM* 43:45–48.

Shampanier, Kristina, Nina Mazar, and Dan Ariely. 2007. "Zero as a Special Price: The True Value of Free Products." *Marketing Science* 26:742–57.

Spence, A. Michael. 1974. *Market Signaling: Informational Transfer in Hiring and Related Processes*. Cambridge, MA: Harvard University Press.

Tenbrunsel, Ann E., and David M. Messick. 1999. "Sanctioning Systems, Decision Frames, and Cooperation." *Administrative Science Quarterly* 44:684–707.

Traulsen, Arne, Torsten Röhl, and Manfred Milinski. 2012. "An Economic Experiment Reveals That Humans Prefer Pool Punishment to Maintain the Commons." *Proceedings of the Royal Society B: Biological Sciences* 279(1743):3716–21.

Tyler, Tom R. 2006. *Why people obey the law*: Princeton University Press. Princeton: NJ

Walker, Henry A., Larry Rogers, and Morris Zelditch Jr. 1988. "Legitimacy and Collective Action: A Research Note." *Social Forces* 67:216–28.

Willer, Robb. 2009. "Groups Reward Individual Sacrifice: The Status Solution to the Collective Action Problem." *American Sociological Review* 74:23–43.

Willer, Robb, Ko Kuwabara, and Michael W. Macy. 2009. "The False Enforcement of Unpopular Norms." *American Journal of Sociology* 115:451–90.

Willer, Robb, Reef Youngreen, Lisa Troyer, and Michael J. Lovaglia. 2012. "How Do the Powerful Attain Status? The Roots of Legitimate Power Inequalities." *Managerial and Decision Economics* 33:355–67.

Xiao, Erte. 2013. "Profit-Seeking Punishment Corrupts Norm Obedience." *Games and Economic Behavior* 77:321–44.

Xiao, Erte, and Fangfang Tan. 2014. "Justification and Legitimate Punishment."

Journal of Institutional and Theoretical Economics 170:168–88.
Yamagishi, Toshio. 1986. "The Provision of a Sanctioning System as a Public Good." *Journal of Personality and Social Psychology* 51:110–16.
Zelditch, Morris. 2001. "Processes of Legitimation: Recent Developments and New Directions." *Social Psychology Quarterly* 64:4–17.
Zelditch, Morris, and Henry A. Walker. 1984. "Legitimacy and the Stability of Authority." *Advances in Group Processes* 1:1–25.
Zelditch, Morris, and Henry A. Walker. 2000. "The Normative Regulation of Power." *Advances in Group Processes* 17:155–78.

BIOS

Ko Kuwabara is an associate professor of organizational behavior at INSEAD Singapore. He received his PhD in sociology from Cornell University. His current research interests, and recent publications, include lay theories of networking (*Academy of Management Review* 2007), cultural differences in social exchange (*Social Psychology Quarterly* 2015; *Psychological Science* 2015), and peer-to-peer markets (*American Journal of Sociology* 2015; *Social Science Research* 2016; *Social Forces* 2017).

Siyu Yu is a PhD student in the Department of Management and Organization at Stern School of Business, New York University. She is broadly interested in social hierarchy and team dynamics. Her current line of work examines how perceptions of social hierarchy affect individuals' organizational outcomes.