Coordination Through Punishment
– An individual Level Perspective –

Felix Albrecht[†], Sebastian Kube[‡,♯]

Preliminary Draft

May 6, 2017

Abstract

JEL-Classification: C9; D03

Keywords: strategy method, conditional contributions, conditional punishment, type classification, public goods game

† University of Marburg; ‡ University of Bonn; ♯ Max Planck Institute for Research on Collective Goods;
* Hertie School of Governance

1 Introduction

- punishment types known from AKT1
- literatur Brandts cooper - other mechanisms
- punishment in coordination games Le Lec
- con1 model -> efficient equilibrium punishment
- paydiff model -> payoff comparison model -> more meaningful in piecewise -> appendix discussion -> lead to coordination on other equilibria
- type distribution
- beta plot
- types predominately stable
- impact on coordination exists
- improves coordination to the efficient equilibrium
- graphs coordination
- improves cooperation ?
- no impact on coordination by FGF CC types
- monte carlo simulation for randomized distribution (see old code)
- Zeiteffekt von Einleitungsfragen

Coordination issues arise routinely in economic circumstances. In microeconomics, the ubiquity of coordination problems within firms, organizations and even industrial branches has been widely acknowledged (e.g., Becker and Murphy, 1992). One prominent game theoretic description of coordination issues is given by the minimum effort game, also known as the weakest-link game: a group member's payoff depends on her own effort (i.e., action) as well as the minimum effort of the group. The higher the minimum effort, the higher is every member's payoff. In contrast to social-dilemma games (e.g., public goods games), any common effort level chosen by all group members is an equilibrium, so it is in no one's interest to deviate upward or downward from the common effort. Hence choosing the most efficient (i.e., payoff- dominant) equilibrium

is a problem of coordination rather than one of cooperation. Many economic and organizational contexts feature situations where agents (e.g., group or team members) must coordinate on a common action with the group's success depending on the least favorable action of a team member. Among canonical examples are teams of assembly-line workers whose overall productivity depends on the least productive member, teams of construction workers whose ability to proceed to the next construction step hinges on every member having completed a task, law firm cases that are only as sound as their weakest part, and even collaboration on scientific projects. Camerer and Knez (1994) have underlined ways weakest-link coordination games can account for within-firm interactions.

Numerous experimental studies have been carried out to determine whether agents are able to collectively coordinate on efficient outcomes. For minimum effort games in particular, ample evidence from various contexts has documented a widespread failure to coordinate on the most efficient, or at least a highly efficient outcome on a long-term basis, starting with the seminal work of Van Huyck, Battalio, and Beil (1990). Various efficiency-enhancing features have been tested experimentally and several of those were found to partly achieve this goal. Among these are smaller groups (Van Huyck, Battalio, and Beil, 1990); higher incentives in the form of ex- ogenous bonuses (Brandts and Cooper, 2007) or lower effort costs (Goeree and Holt, 2005); more refined action space (Van Huyck, Battalio, and Rankin, 2007); commu- nication opportunities including pre-play cheap talk (Blume and Ortmann, 2007); ex post disapproval messages (Dugar, 2010) or centralized communication by a team leader (Brandts and Cooper, 2007); and more homogeneous socio-demographic group composition (Engelmann and Normann, 2010). Except in these very specific set- tings, there appears to be a gradual and pronounced failure to coordinate on the payoff-maximizing equilibrium, even when the stage game is repeated with the same subjects.

We turn to an alternative efficiency-enhancing device, namely voluntary monetary sanctions inflicted on group members who deviate from efficient coordination. Such a mechanism has been established as a powerful force to foster cooperation in public goods games (e.g., Fehr and Gaechter, 2000), suggesting that decentralized, infor- mal sanctions might explain successful cooperation in the field (Ostrom, Walker, and Gardner, 1992). We hypothesize that a similar mechanism may be at work in coordination contexts. For instance, in team projects similar to the examples above, workers may have many opportunities to retaliate against low-effort individuals by not sharing strategic information, refusing future cooperation, and so forth. The sociological literature has long put forth that conventions and norms are often, if not always, enforced by individuals, most of the time in an informal, decentralized, and voluntary manner (Horwitz, 1990). The possibility of sanctions could thus have a strong effect on coordination dynamics as

well as efficiency, potentially explaining high levels of efficient coordination in specific real-world settings.

To examine this hypothesis, we set up an experiment of the minimum effort game based on the work of Van Huyck, Battalio, and Beil (1990). At the beginning of each round, subjects in groups of eight choose an integer effort level between 1 and 7. Then subjects receive anonymous feedback on the effort choices of their fellow group members, and, depending on the treatment, can assign points to them. In the Disapproval treatment, these points simply act as a communication device signal- ing disapproval, with no monetary consequence, as tested by Dugar (2010). In the Punishment treatment, assigning the points imposes a fine on the punished group member, but also comes at a fee to the punisher, with the fine being twice as large as the fee. As in Masclet, Noussair, Tucker, and Villeval (2003), comparing the punishment and disapproval treatments allows us to disentangle what part of the punishment effect is due to implicit ex post communication, e.g., expression of disapproval, and what is due to the monetary consequences of punishment per se. Even though intuitively appealing, it is not straightforward that subjects will use pun- ishment in a coordination game. In contrast to a cooperation game, the individual motivation for punishment is less clear in coordination games where choosing low efforts penalize oneself to a certain extent. This makes it difficult to interpret such choices on the basis of purely selfish or malevolent intentions. Hence reciprocity, which has been found to be a powerful driver of such behavior (Falk, Fehr, and Fischbacher, 2008), does not necessarily lead to punishing behavior in such con- texts. Whether punishment opportunities will be used in this context and whether they increase efficiency – in particular compared to mere disapproval communication opportunities - is the empirical question we aim to shed light on.

To provide an even stronger test, subjects in all treatments first complete eight rounds of play in the baseline minimum effort design without punishment, likely creating a history of low efficiency to be overcome in the next eight rounds with disapproval or punishment opportunities. A similar setup with an initial baseline phase has been used, for instance, by Brandts and Cooper (2006) to study the ef- fect of ex ante communication; Romero (forthcoming) to examine variation in effort cost; and Fatas, Neugebauer, and Perote (2006) to assess the magnitude of a pure 'restart' effect between two successive identical baseline stages. Based on these stud- ies, we expect to find strong path-dependence and, at best, a mild positive restart effect, hence facilitating a strong test of the viability of ex post monetary punish- ment and cheap-talk disapproval as coordination devices. This initial baseline stage distinguishes our Disapproval treatment from an otherwise similar disapproval treat- ment conducted in Dugar (2010).

2 Design and Procedures

The core of our experiment consists of two 10 period repeated public goods games, played in groups of 4 individuals that remain stable for the duration of a game but are randomly rematched between the two. The *RP-game* is a linear public goods game in the tradition of Fehr and Gächter 2000, 2002 that implements punishment on the second stage of individual decision process. Contrary to the RP-game, where the sum of individual contributions determines the size of the public good, the *WP-game* implements the public good as a *minimum effort* game, i.e., the size of public good is determined by the smallest individual effort in a group. Both games implement costly peer punishment in the same way, i.e., for every point subjects invest into reducing a peer's payoff that peer loses three experimental tokens.

To allow us to elicit individual punishment behavior across the two domains, both games implement the punishment stage in the first period as a punishment strategy-method as employed by Kube and Traxler 2011; Albrecht, Kube, and Traxler 2016. The exogenous variation in the contributions in the strategy method allows for elicitation of individual punishment behavior in a repeated game setting in a linear public goods (RP) and weakest link (WP) game.

In a third game we use the by Fischbacher, Gächter, and Fehr 2001 introduced strategy method on the contribution stage in a linear public goods game without punishment to elicit individual conditional contribution types. We relate these contribution types to the individual punishment types and investigate their importance for group success in a repeated public goods and weakest link setting.

2.1 C-Game

The C-game is a standard one-shot linear public-goods game (VCM) with the strategy-method from Fischbacher, Gächter, and Fehr 2001. Subjects are randomly assigned into groups of four. Each subject $i \in \{1, \ldots, 4\}$ is endowed with 20 tokens and decides how many tokens to contribute to the public good, g_i, and how many to keep for herself, $20 - g_i$. Each token allocated to the public good yields a marginal per capita return of 0.4. The payoff function is given by

$$\pi_i^C = 20 - g_i + 0.4 \sum_{j=1}^{4} g_j. \tag{1}$$

Under the assumptions of rational payoff-maximizing behavior, contributing zero is the dominant strategy of the one-shot game. In contrast, the social optimum consists of all players contributing their entire endowment to the public good.

Following the procedure of Fischbacher, Gächter, and Fehr 2001, subjects are first asked to make an unconditional contribution decision, g_i. Using the strategy-method, subjects then make their conditional contribution decisions. They have to indicate their contribution for all 21 possible whole numbers of average contributions among the other group members, $\overline{g}_j := \frac{1}{3}\sum_{j\neq i} g_j$, with $\overline{g}_j \in \{0,1,\ldots,20\}$. After all decisions are made, one group member is randomly drawn. For this subject, the conditional contribution decision is implemented based on the average unconditional contributions of the other three group members. Contributions and payoffs are revealed to the subjects only at the end of the experiment.

2.2 RP-Game

The repeated linear public goods game (RP-game) with costly punishment is implemented in the spirit of Fehr and Gächter 2000, 2002. At the beginning of the game subjects are randomly assigned into groups of four. Each subject $i \in \{1..4\}$ is endowed with 20 tokens and has to decide how many tokens to contribute to the public good, g_i, and how many to keep for himself, $20-g_i$. Each token allocated to the public good yields a marginal per capita return of 0.4 tokens. At the second stage of the game, each subject i can assign punishment points to the other group members $j \neq i$, $d_{ij} \geq 0$. Assigning 1 punishment point costs 1 token for the punisher (1) and reduces the payoff of the punished subject by 3 tokens (2) (e.g., Fehr and Gächter 2002; Herrmann, Thöni, and Gächter 2008). The payoff function is therefore:

$$\pi_i = \underbrace{20 - g_i + 0.4\sum_{j=1}^{4} g_j}_{\text{VCM}} \underbrace{-1\sum_{j\neq i} d_{ij}}_{(1)} \underbrace{-3\sum_{j\neq i} d_{ji}}_{(2)}. \tag{2}$$

Under the assumption of rational selfishness, the unique subgame-perfect Nash equilibrium is zero punishment and zero contributions to the public good.

We innovate on Albrecht, Kube, and Traxler 2016; Kube and Traxler 2011 in that we implement the strategy method at the punishment stage in the first period of the game, rather than playing a one-shot game.[1] Throughout the 10 periods of the experiment subjects make their contribution decisions in the first stage of the game without knowledge of the contribution decisions of their peers in the current period. In the second stage, subjects decide upon deducting points from their

1. The procedure was first applied by Kube and Traxler 2011 as a one-shot implementation and later used by Albrecht, Kube, and Traxler 2016 and **CHINESE**. A similar approach – called 'Conditional Information Lottery (CIL)' – is used in Bardsley 2000. However, the CIL was applied at the contribution rather than the punishment stage. Cheung 2014 uses a strategy method on the punishment stage in a public goods games but reduces the group size to 3 subjects and drastically truncates the range of contribution decisions.

peers based upon the information about the tokens individually submitted to the public good by the other three group members.

The second stage of the first period varies in its setup from the subsequent 9 periods in that it includes the punishment strategy method as originally implemented by Kube and Traxler 2011 (*RPS-game*). In the RPS subjects are confronted with a sequence of contribution triples of the other group members and have to decide on assigning punishment points to other subjects. The details of the procedure are as follows: each subject i faces 11 screens, where each screen presents one contribution triple: $\{g_j^t, g_k^t, g_l^t\}$, with $t \in [1, 11]$; the subindices denote the contributions of the other group members, $i \neq j \neq k \neq l$. One of the 11 triples is given by the actual, 'real' contribution decisions made by the other group members. The remaining ten triples are hypothetical combinations of contributions, each being randomly drawn from a pre-defined set of combinations (see below). In addition, we also randomize the sequence at which individuals face these triples. For each triple, a subject has to decide how many punishment points (if any) to allocate to the other subjects. As the experiment aims at eliciting punishment behavior in a natural way it is important that subjects face contributions from the entire strategy space while at the same time avoiding boredom and overstraining people with too many situations. We therefore partition the strategy space into three intervals: *low (L)*, *intermediate (M)*, and *high (H)* contributions with $g^L \in \{0, ..., 4\}$, $g^M \in \{5, ..., 15\}$, $g^H \in \{16, ..., 20\}$. Based on these partitions, we considered 10 combinations of low, intermediate and high contributions:

Within each of the 10 hypothetical contribution combinations, we randomly draw from a set of 8

Table 1: Composition of Contribution Triplets

Hypothetical:

$\{g^L, g^L, g^L\}$	$\{g^L, g^L, g^M\}$	$\{g^L, g^L, g^H\}$	$\{g^L, g^M, g^M\}$	$\{g^L, g^M, g^H\}$
$\{g^L, g^H, g^H\}$	$\{g^M, g^M, g^M\}$	$\{g^M, g^M, g^H\}$	$\{g^M, g^H, g^H\}$	$\{g^H, g^H, g^H\}$

+ Real: $\{\mathbf{g_j}, \mathbf{g_k}, \mathbf{g_l}\}$

different triples (the triples are reported in the Online Appendix A). A subject could then face, for instance, $\{0, 2, 3\}$ for the combination $\{g^L, g^L, g^L\}$ and $\{1, 2, 10\}$ for $\{g^L, g^L, g^M\}$, etc. A different subject might face $\{1, 3, 3\}$ for the former and $\{0, 2, 14\}$ for the later.[2]

All subjects know that 10 out of the 11 contribution triples are hypothetical. It is also common knowledge that only the punishment decisions for the real contribution triple is payoff relevant.

2. If, by chance, the triple $\{0, 2, 14\}$ would correspond to the real combination of contributions, the subject would not face this triple. Instead a different triple from the pre-defined set of contribution triples for $\{g^L, g^L, g^M\}$ would be randomly drawn.

However, subjects neither know which one is the 'real' triple, nor do they know the procedure to generate the hypothetical triples. Following this protocol, we observe 3×11 punishment decisions for each subject. As discussed below, our analysis will explore only the choices made for the 30 hypothetical contributions.

Once subjects complete their punishment decisions for the 11 screens they are informed about the payoffs for period 1 and continue to period 2. Subjects continue playing the public goods game for the whole duration of the RP-game. In the subsequent period 2 to 10, however, subject do not face hypothetical triples but are only presented with the real contributions of the other subjects. We made clear to subjects that the group composition in all 10 periods would remain constant and that they would interact repeatedly.

2.3 WP-Game

The structure of the second game (*WP-game*) is similar to the RP-game in that subjects play repeatedly for 10 periods in fixed groups of four, contribute to a public good on the first stage of the game, are allowed to sanction their peers on the second stage of the game, and face a second stage strategy method in the first period (*WPS-game*). We also keep the endowments (20 tokens) and the punishment technology (1 point assigned, reduces the punishee's payoff by 3 tokens) constant in respect to the RP-game. However, we vary the payoff function in that the size of the public good in the WP-game is now determined by the smallest individual contribution rather than the sum of all contributions. The individual payoff function for this *weakest link* game is presented in equation 3.

$$\pi_i = \underbrace{20 - g_i + 1.6 \times \min_{i,j,k,l}(g_j)}_{\text{Weakest Link}} \underbrace{-1 \sum_{j \neq i} d_{ij} - 3 \sum_{j \neq i} d_{ji}}_{\text{Punishment}}. \tag{3}$$

The weakest link game differs from a linear public goods game in respect to its incentives and Nash-equilibria. Subject in a linear public goods game have the selfish incentive to free ride on others by contributing less than them, making *zero* contribution the only Nash-equilibrium in pure strategies. Contrarily, in a weakest link setting every common effort level chosen by all members of a group ($g_i = g_j = g_k = g_l$) are Nash-equilibria. In our WP-game all 21 equilibria are pareto-ranked with $g_{i,j,k,l} = 20$ being the *most efficient (payoff dominant)* and $g_{i,j,k,l} = 0$ being the *risk dominant* equilibrium.

For the implementation of the strategy method WPS-game hypothetical triples were drawn from the predefined contribution triple space that was employed for the RPS-game. Triples were

again drawn from the sets at random and shuffled on an individual level before the treatment commenced.

Felix: Wie viele sahen gleiche triple?

2.4 Implementation

We evaluate data for 228 subjects collected in 10 sessions in the *BonnEconLab* at the University of Bonn, Germany. For every subject we observe 2×30 punishment decisions with exogenous contribution variation in the strategy method implementation of the *RPS* respectively *WPS-game* and 21 contribution decisions from the *C-game*. In addition we observe contribution and punishment behavior that is independent on the group level for 57 groups of four over the 10 periods of the *RP* and *WP-game* (2×10 observations).

Because the RP and WP-game only differ in their payoff functions confusion among the subjects were of major concern to us. We took great care to insure thorough understanding of the treatment differences by subjects. Before each treatment subjects had to correctly answer a set of control questions that specifically tested comprehension of payoff calculations. We further differentiated the treatments by using treatment specific language. For transfer to the public good in the *RP-game* we used the German term for *'contribution'*, for transfer to the public good in the *WP-game* we used the German word for *'effort'*[3] and emphasized at the beginning of the subsequent treatment that we intentionally changed the terminology to avoid confusion. Lastly for the WP-game subjects received a printed payoff matrix together with the printed instructions, which was not available for the RP-game, to also visually differentiate the two treatments.[4]

The experiment was conducted using the experimental software *ztree* (Fischbacher 2007). Experimental subjects were recruited from the BonnEconLab's subject pool using Hroot [CITE]. Standard experimental procedures were followed. Subjects were randomly seated, communication during sessions was strictly forbidden and would have led to exclusion from the experiment. Experimental groups were randomly rematched between treatments but remained constant during treatments. Subjects were informed about the payoffs from the C-game only at the end of the experimental session. For the RP and WP-game subjects received information on payoffs after every period. Including a follow-up questionnaire a session lasted approximately 140 minutes with subjects earning on average [PAYOFF] including a 5 Euro show-up fee.

3. Specifically we used the word 'Beitrag' for 'contribution' and 'Aufwendung' for 'effort' and their respective verbs to express the act of transferring tokens to the public good.

4. Section **??** in supporting online material presents the payoff matrix translated into English.

3 Punishment Type Distributions

We follow Albrecht, Kube, and Traxler 2016 in the classification of punishment types by classifying subjects' punishment behavior with respect to their sanctioning behavior towards others' deviation from full contribution (effort).[5] For each of the 228 individuals we estimate model 4 for the 30 punishment observations obtained from the hypothetical contributions in the RPS(WPS)-game.

$$d_{ij} = \alpha_i + \beta_i(20 - g_j) + \varepsilon_i. \tag{4}$$

Like in Albrecht, Kube, and Traxler 2016 subjects are classified into three behavioral categories, i.e., 'Non-Punisher','Pro-social Punisher', and 'Anti-social Punisher'. A fourth category 'Non-classifiable' captures subjects whose punishment pattern is not captured by the definition of the previous three classes. The three behavioral categories are defined in the following ways:

1. A subject is classified as a 'Non-Punisher' (*NPun*) if she assigns zero punishment points in all of the 30 punishment decisions, i.e., $d_{ij} = 0$ for all g_j. In equation (4), this is depicted by $\hat{\alpha}_i = \hat{\beta}_i = 0$.

2. Subjects that target their punishment towards those that contribute little or nothing to the public good have a punishment pattern that is upward sloping in $(20 - g_i)$. These subjects, with $\hat{\beta}_i > 0$ and $p \leq 0.01$, are classified as pro-social punishers (*Pun*).

3. Subjects are classified as anti-social punishers (*APun*), if their punishment is either increasing in the other's contribution g_j, i.e., if $\hat{\beta}_i < 0$ and $p \leq 0.01$, or if they display a significant positive but unsystematic level of punishment: $\hat{\alpha}_i > 0$ with $p \leq 0.01$ and an insignificant slope coefficient $\hat{\beta}_i$ with $p > 0.01$.[6]

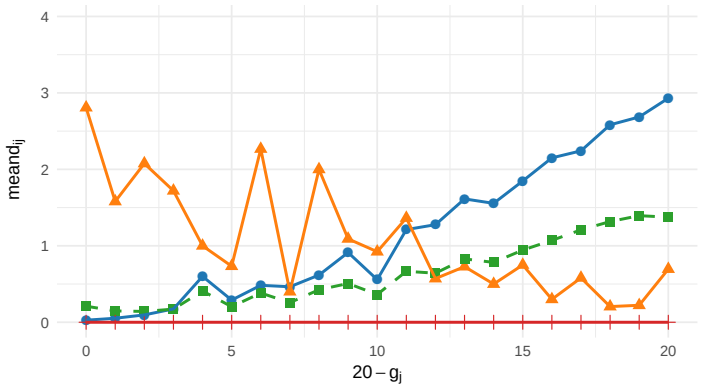
3.1 Public Goods Game Punishment Types

We begin by replicating Albrecht, Kube, and Traxler 2016, classifying individual level punishment patterns in a linear public goods game (RPS) with punishment strategy method. Figure 1 presents the distribution of punishment patterns for our 228 experimental subjects. 48.7% of our subjects

5. Section C in the supporting online material discusses why we employ others' contributions g_j rather than others' payoff pre-punishment to estimate punishment behavior.

6. The literature typically defines anti-social punishment in reference to a subject's own contribution, i.e., if the punishment-receiving subject contributed a larger or equal amount to the public good compared to the punishing individual (e.g., Herrmann, Thöni, and Gächter 2008). Since our classification does not consider a punisher's own contribution g_i, it deviates from this self-centered notion of anti-social punishment. It nevertheless captures patterns of punishment that is targeted towards high contributors.

Figure 1: RPS-game Punishment Types



Type	N	%
Pun	111	48.7
NPun	88	38.6
APun	13	5.7
NCL	16	7.0
Total	228	100.0

Notes: Punishment type distribution and average punishment patterns (in the $20 - g_j$-space) in the RPS-game for the different types: pro-social punishers (*Pun*), non-punishers (*NPun*), anti-social punishers (*APun*), and non-classified punishment profiles (*NCL*). To ease illustration, the pattern for the latter is not plotted.

show pro-social punishment patterns, punishing low contributors more heavily than high contributors. 38.6% of subjects do not invest in peer-punishment in any of the 30 decision situations. 5.7% punish anti-socially deducting more points from high contributors than low contributors. In 7% of the cases subjects' behavior didn't meet one of the three classifications and remained unclassified.

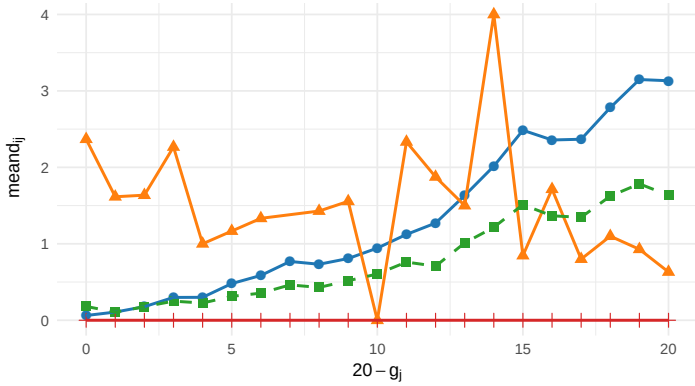
3.2 Weakest Link Game Punishment Types

In a next step we classify individual level sanctioning behavior in a weakest link game setting. As documented by **Lec2015** subjects engage in costly peer-sanctions in coordination game environments as a means to foster coordination. In line with their findings, we too observe considerable sanctions in the weakest link game setting (WPS). Figure 2 shows the distribution of types and their respective average sanctioning behavior in the WPS-game. We observe an increase in pro-socially sanctioning *Pun*-types compared to the public goods setting. 53.1% of subjects sanction peers displaying low effort levels more strongly than if they display larger efforts. We further observe a reduction in non-sanctioning *NPun* (30.3%) and anti-socially sanctioning *APun* (3.1%) individuals. Finally we observe an increase in non-classifiable individuals (13.6%).

3.3 Individual Cross-Domain Punishment Behavior

Combining the two punishment classifications across the two domains allows us to elicit the *individual punishment type stability*. Figure 3 shows the results. The majority (67.7%) of subjects show a consistent punishment type across the two domains. This includes Pun, NPun, and APun

Figure 2: WPS-game Punishment Types



Type	N	%
Pun	121	53.1
NPun	69	30.3
APun	7	3.1
NCL	31	13.6
Total	228	100.0

Notes: Punishment type distribution and average punishment patterns (in the $20 - g_j$-space) in the RPS-game for the different types: pro-social punishers (*Pun*), non-punishers (*NPun*), anti-social punishers (*APun*), and non-classified punishment profiles (*NCL*). To ease illustration, the pattern for the latter is not plotted.

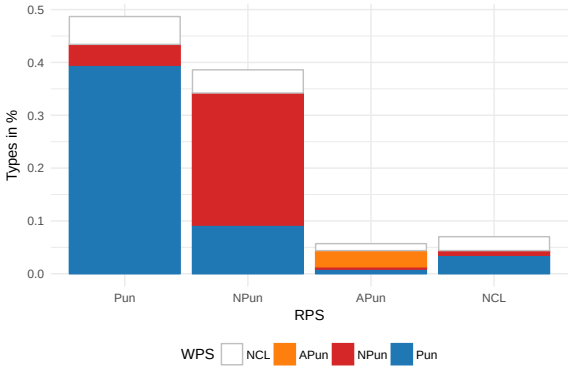
types. Among the switchers the majority of subjects appears to be changing their punishment behavior to be more pro-social. 21 subjects that exert no punishment in the RP-game display a pro-social pattern in the WPS-game.

4 Joint Punishment Demand

It remains to be seen whether this individual behavioral change translates into differences in global punishment demand in the two settings (RPS and WPS). Table 2 presents the results for pseudo panel regressions with individual level fixed effects for the two strategy method settings. Game and screen order are used as time variance to construct the pseudo panel and capture potential ordering effects. As figure 1 and figure 2 indicated we find significant positive punishment demands for deviations from fully contributing one's endowment (highest potential effort level) for both settings (RPS and WPS) in the respective estimations in column 1 and 2. The demand for sanctions in the WPS game appears hereby to be larger than the demand for punishment in the RPS game. The estimation in column 3 supports this. The interaction effect $D.WPS \times (20 - g_j)$ in column 3 is positive and significant on the 5 percent level indicating harsher sanctions for every token kept in the private account in the WPS game over the RPS game.

However, this does not reveal from where these additional sanctions stem. As stated above despite remarkable punishment type stability within the experimental subjects we still observe considerable behavioral fluctuations between the two settings. We therefore investigate whether the additional demand for sanctions is an expression of all subjects adapting their sanctioning behavior to the new incentives or just a sub-population, and if this behavioral adaptation is po-

Figure 3: Type Distribution RPS- and WPS-game



	WPS-game				
RPS	Pun	NPun	APun	NCL	Total
Pun	90	9	0	12	111
%	[39.5]	[3.9]	[0.0]	[5.3]	[48.7]
NPun	21	57	0	10	88
%	[9.2]	[25.0]	[0.0]	[4.4]	[38.6]
APun	2	1	7	3	13
%	[0.9]	[0.4]	[3.1]	[1.3]	[5.7]
NCL	8	2	0	6	16
%	[3.5]	[0.9]	[0.0]	[2.6]	[7.0]
Total	121	69	7	31	228
%	[53.1]	[30.3]	[3.1]	[13.6]	[100.0]

Note: More than 65% of subjects remain consistent across these two games in their punishment behavior.

Table 2: Punishment Demand Across Games

	Assigned Punishment d_{ij}			
	RPS	WPS	Joint	Stable
	(1)	(2)	(3)	(4)
$(20-g_j)$	0.068***	0.085***	0.068***	0.086***
	(0.007)	(0.007)	(0.007)	(0.009)
$D.WPS$			−0.017	0.030
			(0.031)	(0.029)
$D.WPS \times (20-g_j)$			0.017**	0.001
			(0.007)	(0.006)
Intercept	−0.005	−0.019	−0.000	−0.055
	(0.067)	(0.070)	(0.062)	(0.083)
Observations	6,840	6,840	13,680	9,240
adj.R^2	0.222	0.251	0.201	0.263
AIC	18,036	19,974	41,357	26,656
BIC	18,043	19,980	41,379	26,677

tentially already fully reflected in the two-domain punishment classification.

Column 4 in table 4 presents the results for a pooled RPS and WPS estimation reduced to the 154 *type-stable* subjects (Pun, NPun, APun) along the main diagonal excluding NCL. The demand for punishment $(20-g_j)$ increases slightly with respect to column 3 and can be explained with the fact that the majority of stable individuals are Pun-type subjects that generate the bulk of positive punishment demand. More interestingly, none of the other coefficients remain significant on any conventional level. This indicates that type-stable subjects are *not* the driver of the increased punishment demand in the WPS game, meaning that the increased demand for sanctions are a result of individuals that change their sanctioning behavior across the two domains.

5 Contribution × Punishment Types

In the previous sections we studied the distribution of punishment types across a linear public goods game and a weakest link game. We will now continue to cross-link these punishment classifications with individually elicited contribution types.

5.1 Contribution Types

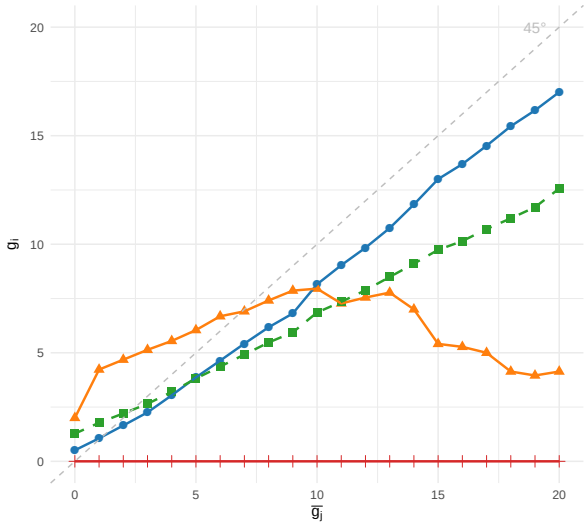
In similar fashion to Albrecht, Kube, and Traxler 2016 we use the C-game observations to classify contribution types in the tradition of Fischbacher, Gächter, and Fehr 2001. We follow Albrecht, Kube, and Traxler 2016 in that we too classify subjects using linear regression results, deviating from the original Fischbacher, Gächter, and Fehr 2001, who use Spearman rank correlation. In the C-game subjects had to decide upon their *conditional contribution* decision with respect to 21 displayed possible group mean contributions. For each individual we estimate equation 5 in which the average contribution of the other three group member $\overline{g}_j$ explains the individual contribution decision g_i.

$$g_i = a_i + b_i \overline{g}_j + e_i \tag{5}$$

Subjects are then classified in accordance with **??**, i.e., *Conditional Contributors* (CC) show a positive slope coefficient b_i, significant on the 1 percent level, *Free Rider* (FR) do not make positive contributions in any of the 21 decision situations, *Triangular Contributors* (TC) show a pattern that initially increases but starts to decrease at a certain point.[7] Similar to the the punishment classification an additional group is introduced to capture behavioral patterns that do not fit with any of

7. As in **??** and **??** we classify TC types via eyeballing.

Figure 4: C-game Contribution Types



Type	N	%
CC	149	65.4
FR	42	18.4
TC	22	9.6
NC	15	6.6
Total	228	100.0

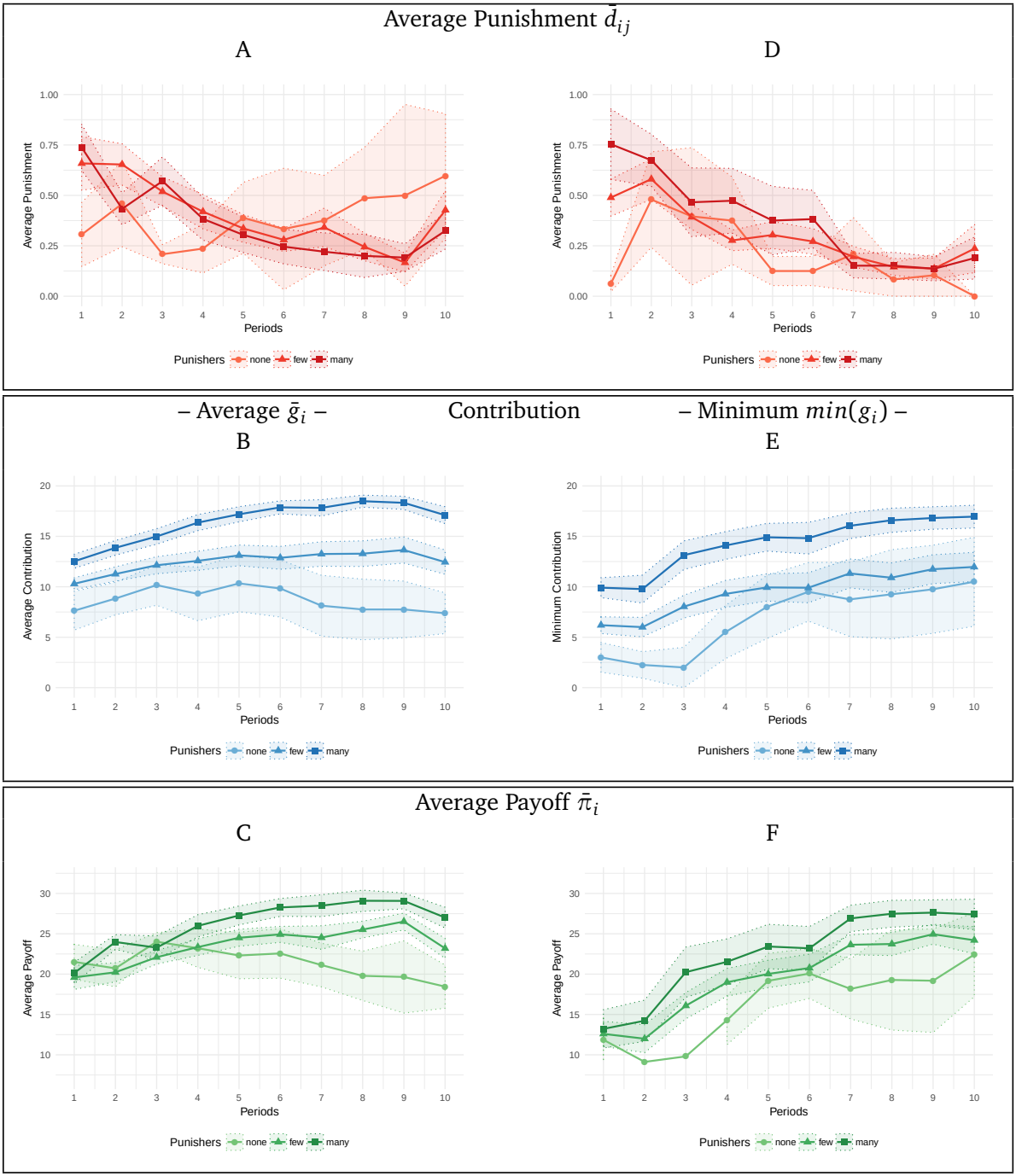
Notes: The figure presents the distribution of contribution types, following Fischbacher, Gächter, and Fehr 2001 and Fischbacher and Gächter 2010, and the average cooperation patterns for the different types: *Conditional Cooperators (CC)*, *Free-Riders (FR)*, *Triangular Contributors (TC)*, and *Non-classified (NC)* cooperation patterns. To ease illustration, the pattern for the latter is not plotted.

the previous groups (NC). Figure 4 presents the results for the contribution type classification. A large majority of subjects (65.4%) show upward sloping contribution patterns and are classified as conditional contributors. Another 18.4% of individuals do not transfer positive amounts to the public goods for any of the 21 group contributions means. Another 9.6% of subjects are classified via eyeballing as triangular contributors. Finally 6.6% individuals did no fit with any of the previous 3 classifications and were labeled an non-classifieds.

6 Impact of Type Prevalence on Group Outcomes

To quickly recap we found that individuals are heterogeneous in their punishment behavior. The majority of subjects, when classified with a fairly simple approach, remain type-stable across two public goods games with differing incentive structures. We further find that the average demand for sanctions within the weakest link setting (WPS) is significantly larger than in the linear public goods setting (RPS) and that this increased demand stems from the non-type-stable individuals, who on average increase their demand for sanctions.

Figure 5: Average Group Outcomes by Type Prevalence

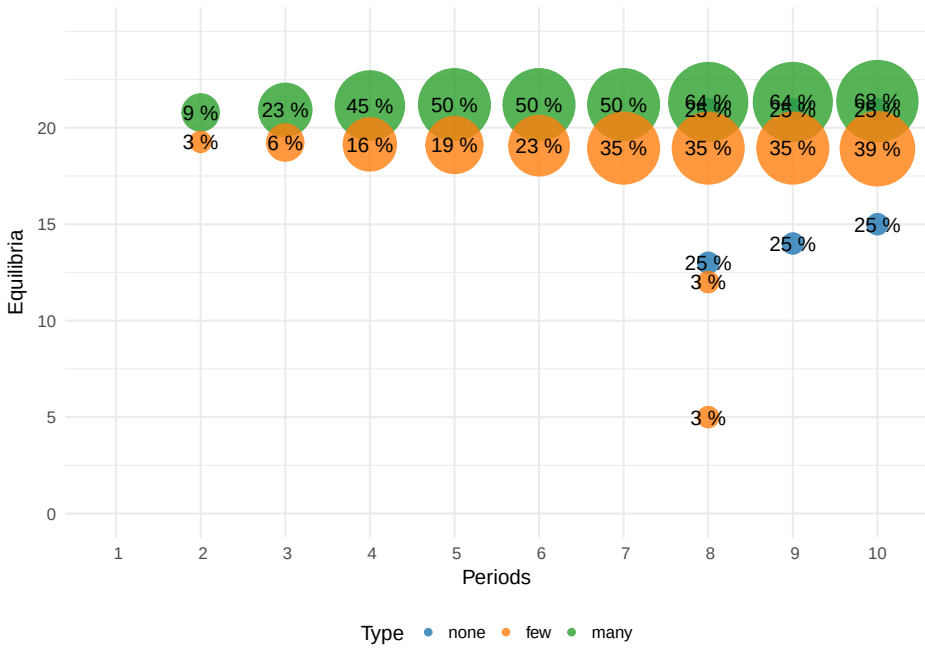


Note: More than 65% of subjects remain consistent across these two games in their punishment behavior.

6.1 Contributions

6.2 Coordination

Figure 6: Distribution of Equilibrium Coordination by Type Prevalence



Note: Fraction of successful equilibrium coordination by the 4 groups *without* Pun-types, 33 with *few*, and 20 groups with *many* Pun-type subjects, as classified based on WPS behavior. Groups *without* Pun-type subjects fail to coordinate on any of the 21 possible equilibria in any of the 10 periods. Groups with *few* Pun-types coordinate 5 times on equilibria that are not the most efficient (payoff-dominant) equilibrium. In the majority of cases groups with *few* Pun-types coordinate on the payoff-dominant equilibrium when successfully coordinating. Groups with *many* Pun-types exclusively coordinate successfully on the payoff dominant equilibrium. All groups fail to tacitly coordinate on an equilibrium in the first period.

References

Albrecht, Felix, Sebastian Kube, and Christian Traxler. 2016. “Cooperation and Punishment: The Individual-Level Perspective.”

Bardsley, Nicholas. 2000. “Control Without Deception : Individual Behaviour in Free-Riding Experiments Revisited.” *Experimental Economics* 3:215–240.

Cheung, Stephen L. 2014. “New insights into conditional cooperation and punishment from a strategy method experiment.” *Experimental Economics* 17 (1): 129–153.

Fehr, Ernst, and Simon Gächter. 2000. “Cooperation and Punishment in Public Goods Experiments.” *American Economic Review,* Working Paper, 90 (4): 980–994.

———. 2002. “Altruistic punishment in humans.” *Nature* 415 (6868): 137–40.

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

Fischbacher, Urs, and Simon Gächter. 2010. “Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments.” *American Economic Review* 100 (1): 541–556.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2001. “Are people conditionally cooperative? Evidence from a public goods experiment.” *Economics Letters* 71 (3): 397–404.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. “Antisocial punishment across societies.” *Science* 319 (5868): 1362–7.

Kube, Sebastian, and Christian Traxler. 2011. “The Interaction of Legal and Social Norm Enforcement.” *Journal of Public Economic Theory* 13 (5): 639–660.

Supporting Online Material

(*For Online Publication Only*)

Coordination Through Punishment – An individual Level Perspective –

Felix Albrecht[†], Sebastian Kube[‡,♯]

May 6, 2017

† University of Marburg; ‡ University of Bonn; ♯ Max Planck Institute for Research on Collective Goods; * Hertie School of Governance

Contents

List of Tables

List of Figures

A Contribution Triplets

Below we list the contribution triples that were used within each combination of g^L, g^M and g^H (see Table 1). Before the experiment, these 10 × 8 triples were randomly generated by sampling with replacement from the corresponding sets g^L, g^M, g^H. Each player then faced a randomly selected triple within each combination 1 – 10. If the selected triple would by chance correspond to the real triple, the subject would not face this situation but instead another one of the pre-defined contribution triples for the corresponding combination.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	(g^L, g^L, g^L):	(0,0,0)	(0,2,3)	(1,1,3)	(1,2,2)	(1,2,3)	(1,2,4)	(1,3,3)	(1,3,4)
(2)	(g^L, g^L, g^M):	(0,1,5)	(0,2,8)	(0,2,14)	(1,2,10)	(1,2,12)	(1,3,14)	(2,2,6)	(2,3,12)
(3)	(g^L, g^L, g^H):	(0,3,18)	(1,2,20)	(1,3,19)	(1,4,20)	(2,2,18)	(2,2,19)	(3,3,18)	(4,4,17)
(4)	(g^L, g^M, g^M):	(0,9,11)	(0,5,12)	(0,13,14)	(1,10,15)	(2,6,8)	(2,9,11)	(2,10,15)	(3,13,14)
(5)	(g^L, g^M, g^H):	(0,6,19)	(0,14,17)	(2,6,17)	(2,8,20)	(2,11,19)	(3,7,18)	(4,8,17)	(4,10,20)
(6)	(g^L, g^H, g^H):	(0,18,19)	(1,19,19)	(2,18,19)	(2,18,20)	(2,19,19)	(3,18,20)	(3,19,19)	(4,19,20)
(7)	(g^M, g^M, g^M):	(5,7,12)	(5,14,16)	(6,6,9)	(6,10,10)	(7,8,9)	(7,10,13)	(7,14,16)	(8,9,11)
(8)	(g^M, g^M, g^H):	(5,5,17)	(5,8,18)	(6,11,20)	(8,15,17)	(9,12,18)	(9,15,18)	(11,15,19)	(12,15,19)
(9)	(g^M, g^H, g^H):	(5,18,20)	(7,18,19)	(9,18,20)	(11,17,17)	(12,17,18)	(12,18,18)	(14,17,20)	(15,17,19)
(10)	(g^H, g^H, g^H):	(17,17,19)	(17,18,19)	(17,18,20)	(17,19,19)	(17,19,20)	(18,18,19)	(18,18,20)	(20,20,20)

B Instructions

Below you find the English set of instructions used at FSU. The German and Japanese set of instructions can be requested from the authors. The first part describes a public goods game without punishment in the tradition of (Fischbacher, Gächter, and Fehr 2001), which was played ahead of the core games with punishment. The second part describes the *One-Shot* game, the third the *Repeated Interaction* game respectively.

C Contribution vs. Payoff Discussion

Table 1: Contribution versus Payoff as Explanatory Variables

	Individually Assigned Punishment d_{ij}			
	RPS		WPS	
	(1)	(2)	(3)	(4)
$(20 - g_j)$	0.068***		0.085***	
	(0.007)		(0.007)	
π_j		0.062***		0.038***
		(0.006)		(0.003)
Intercept	−0.005	−0.962***	−0.019	0.182***
	(0.067)	(0.158)	(0.070)	(0.057)
Observations	6,840	6,840	6,840	6,840
Adjusted R^2	0.222	0.098	0.251	0.048
AIC	18,036	19,054	19,974	21,619
BIC	18,043	19,061	19,980	21,626

Note: Pseudo panel estimation with individual level fixed effects for 228 subjects. Cluster robust standard errors in parentheses. P-values: 0.1 / 0.05 / 0.01 : * / ** / ***. Column (1), (3), and (5) only RP-game, (2), (4), and (6) only WP-game observations with 30 observations per subjects in each game.

Table 2: Contribution Difference vs. Payoff Difference as Explanatory Variables

	Individually Assigned Punishment d_{ij}		
	(1)	(2)	
$max(\pi_j - \pi_i, 0)$	−0.008	−0.008	$max(g_j - g_i, 0)$
	(0.006)	(0.006)	
$max(\pi_i - \pi_j, 0)$	0.120***	0.120***	$max(g_i - g_j, 0)$
	(0.012)	(0.012)	
$D.WPS \times max(\pi_j - \pi_i, 0)$	−0.003	−0.003	$D.WPS \times max(g_j - g_i, 0)$
	(0.006)	(0.006)	
$D.WPS \times max(\pi_i - \pi_j, 0)$	−0.004	−0.004	$D.WPS \times max(g_i - g_j, 0)$
	(0.012)	(0.012)	
$D.min(g_{jkl})$	0.278***	0.278***	$D.min(g_{jkl})$
	(0.057)	(0.057)	
$D.WPS \times D.min(g_{jkl})$	0.043	0.043	$D.WPS \times D.min(g_{jkl})$
	(0.072)	(0.072)	
Intercept	0.061	0.061	Intercept
	(0.056)	(0.056)	
Observations	13,680	13,680	Observations
Adjusted R^2	0.319	0.319	Adjusted R^2
AIC	39,175	39,175	*AIC*
BIC	39,220	39,220	*BIC*

Note: Pseudo panel estimation with individual level fixed effects for 228 subjects. Cluster robust standard errors in parentheses. P-values: 0.1 / 0.05 / 0.01 : * / ** / ***. Column (1), (3), and (5) only RP-game, (2), (4), and (6) only WP-game observations with 30 observations per subjects in each game.

D Additional Analyses

Table 3: Type Dristribution RP- and WP-game

	t_sdiff_61								
t_sdiff_bz41	-12	-10	-1	0	2	20	23	100	Total
-11 No.	0	1	0	0	0	0	0	0	1
%	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.4
-10 No.	0	50	0	6	0	0	0	6	62
%	0.0	21.9	0.0	2.6	0.0	0.0	0.0	2.6	27.2
-1 No.	0	0	38	3	0	0	0	5	46
%	0.0	0.0	16.7	1.3	0.0	0.0	0.0	2.2	20.2
0 No.	0	8	14	57	0	0	0	9	88
%	0.0	3.5	6.1	25.0	0.0	0.0	0.0	3.9	38.6
2 No.	0	0	0	1	1	0	0	1	3
%	0.0	0.0	0.0	0.4	0.4	0.0	0.0	0.4	1.3
20 No.	1	0	0	0	0	2	0	3	6
%	0.4	0.0	0.0	0.0	0.0	0.9	0.0	1.3	2.6
100 No.	0	7	2	2	0	0	1	10	22
%	0.0	3.1	0.9	0.9	0.0	0.0	0.4	4.4	9.6
Total No.	1	66	54	69	1	2	1	34	228
%	0.4	28.9	23.7	30.3	0.4	0.9	0.4	14.9	100.0

Note:

D.1 Payoff Comparison

Table 4: Punishment Demand

	Assigned Punishment d_{ij}							
	RPS	WPS	Joint	stable	RPS	WPS	Joint	stable
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$(20-g_j)$	0.068***	0.085***	0.068***	0.086***	0.056***	0.072***	0.052***	0.073***
	(0.007)	(0.007)	(0.007)	(0.009)	(0.007)	(0.007)	(0.007)	(0.009)
$D.WPS$			−0.017	0.030			−0.025	0.026
			(0.031)	(0.029)			(0.031)	(0.030)
$D.WPS \times (20-g_j)$			0.017**	0.001			0.019***	0.001
			(0.007)	(0.006)			(0.007)	(0.006)
$D.min(g_{jkl})$					0.473***	0.424***	0.580***	0.450***
					(0.051)	(0.045)	(0.064)	(0.066)
$D.WPS \times D.min(g_{jkl})$							−0.130*	−0.030
							(0.076)	(0.078)
Intercept	−0.005	−0.019	−0.000	−0.055	0.025	0.009	0.036	−0.027
	(0.067)	(0.070)	(0.062)	(0.083)	(0.065)	(0.069)	(0.061)	(0.082)
Observations	6,840	6,840	1,3680	9,240	6,840	6,840	1,3680	9,240
adj.R^2	0.222	0.251	0.201	0.263	0.246	0.267	0.223	0.280
AIC	18,036	19,974	41,357	26,656	17,821	19,827	40,978	26,442
BIC	18,043	19,980	41,379	26,677	17,835	19,841	41,016	26,477

Table 5: Mean Contribution g_i in RP- and WP-game

	μ_{g_i}	SE
RP	10.798	(0.457)
WP	13.447	(0.397)
t-test		0.000

Note: