No. 104

Irenaeus Wolff
Bettina Rockenbach

Designing Institutions for Social Dilemmas

Research Paper Series
Thurgau Institute of Economics and Department of Economics
at the University of Konstanz

Designing Institutions for Social Dilemmas

Bettina Rockenbach and Irenaeus Wolff

University of Cologne and Thurgau Institute of Economics/University of Konstanz

March 2016

German Economic Review, forthcoming

Abstract:
Considerable experimental evidence has been collected on rules enhancing contributions in public goods dilemmas. These studies either confront subjects with pre-specified rules or have subjects choose between different rule environments. In this paper, we completely endogenize the institution design process by asking subjects to design and repeatedly improve rule sets for a public goods problem in order to investigate which rules social planners facing a social dilemma "invent" and how these rules develop over time. We make several noteworthy observations, in particular the strong and successful use of framing, the concealment of individual contribution information and the decreasing use of punishment.

Keywords: Public good; strategy method; experiment; public choice

JEL-Classification: C72, C92, D71, D72

Addresses:

BR: University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany,
bettina.rockenbach@uni-koeln.de
IW: Thurgau Institute of Economics (TWI), Hauptstrasse 90, 8280 Kreuzlingen, Switzerland,
wolff@twi-kreuzlingen.ch

Acknowledgements:

We are deeply indebted to Reinhard Selten for proposing this research and for our intensive discussions throughout the project. Special thanks go to our seminar participants for their enthusiasm while taking part in this study. We also thank participants at seminars in Bonn and Erfurt for their inspiring comments, as well as our colleagues at the chair of microeconomics for their tireless help in testing the experimental software programs, as well as for the fruitful discussions along the way.

I. Introduction

Understanding the determinants of human cooperation is one of the most challenging questions in economics. Human cooperation in social dilemmas is particularly puzzling because the conflict between individual and collective interests creates the well-known free-rider problem.[1] Repeated interactions in controlled social dilemma experiments show that initially positive cooperation levels decline over time and thus nurture Hardin's pessimistic vision.[2]

Yet, in order to overcome the breakdown of cooperation and to disentangle different motives for cooperation and defection, researchers have incorporated a huge variety of institutional regulations into the workhorse for studying social dilemmas, the experimental public goods provision game. Among the most prominent regulations examined are punishment opportunities,[3] communication,[4] leadership,[5] reputation opportunities,[6] and ostracism.[7] In this "first generation" of experiments, subjects face experimenter-determined institutional rules. This yields valuable information on the performance of these rules for fostering cooperation, but cannot answer the question of whether subjects would actually select a certain rule when having the choice. Consequently, subsequent studies have addressed the question of institutional choice among (mostly two) pre-specified rules, both by ballot voting[8] and by 'voting-with-one's-feet'[9]. These "second generation" studies provide important insights into the institutional rules subjects choose initially as well as in the longer run when different (experimenter-given) rules are at choice. However, are these the rules social planners facing a social dilemma would develop themselves? Which are the rules planners would "invent" and how would these rules develop over time in the light of their ability to overcome the social dilemma? An experimental answer to these questions requires to not only endogenize the choice process, but the entire institution formation process. This is exactly what this paper aims at.

In this paper, we introduce an experiment of a "third generation" in which the institution design process is entirely endogenous. The subjects in our experiment act as rule makers empowered to shape the institutional environment of a public goods provision game. They are free in "inventing"

[1] Hardin (1968), Ostrom (1998).

[2] Davis and Holt (1993), Ledyard (1995), Ostrom (2000), Fischbacher and Gächter (2010).

[3] E.g. Yamagishi (1986), Fehr and Gächter (2000), Masclet, Noussair, Tucker, and Villeval (2003), Nikiforakis (2008), Nikiforakis and Normann (2008), or Falk, Fehr, and Fischbacher (2005).

[4] E.g. Isaac and Walker (1988), Ostrom, Gardner and Walker (1994), Cason and Khan (1999), Brosig, Weimann and Ockenfels (2003), or Bochet, Page, and Putterman (2006).

[5] E.g. Vesterlund (2003), Potters, Sefton and Vesterlund (2005), Arbak and Villeval (2013), or Güth, et al. (2007).

[6] E.g. Milinski, et al. (2006), or Sommerfeld, et al. (2007).

[7] E.g. Cinyabuguma, Page and Putterman (2005), Maier-Rigaud, Martinsson and Staffiero (2005), or Güth, et al. (2007).

[8] E.g. Potters, Sefton and Vesterlund (2005), Sutter, Haigner and Kocher (2005), Guillen, Schwieren and Staffiero (2007).

[9] E.g. Gürerk, Irlenbusch and Rockenbach (2006 and 2014), Rockenbach and Milinski (2006), Kosfeld, Okada and Riedl (2009).

any institutional setting and not bound to any pre-specified rules, as long as they do not change the social-dilemma incentive structure. The subjects' incentive for acting as "good" rule makers is that their payment strongly depends on the social benefits that game playing students accrue under the designed rule. Based on performance feedback, subjects may improve their institutional design. To equip the institution designers with a profound understanding and experience in the public goods provision game and an intense experience in the institution design process, we followed Reinhard Selten's advice and conducted the experiment as a *strategy seminar* (Selten 1967).

In his seminal paper, Selten (1967) introduced an experimental method, called the *strategy method* and applied it to oligopoly situations.[10] The idea of the strategy method is that subjects, which are highly familiar with the game situation through repeated play of the game, develop strategies for acting in the game. In a series of tournaments, subjects learn about the performance of their strategy and have the possibility to improve the strategy based on this feedback. Usually the strategy method is applied within the framework of a "strategy seminar", a students' seminar over several months (a semester) to provide a sufficient time frame.[11] The obvious advantage of the strategy method is that highly experienced subjects develop strategies that they continuously improve based on performance feedback. The apparent downside is that conducting the experiment is very time consuming, so that the research focus is not on collecting a sufficient number of independent observations for statistical analysis, but rather on improving our understanding on how highly experienced experts solve (complex) decision problems. Since the high level of complexity also limits us in gathering observations sufficient for sensible statistical analysis, our study shares the explorative nature of the previous applications of the strategy method.

Our experiment provides us with remarkable insights, which we extensively discuss in the final section. Most noteworthy seems to be: (1) institution designers make extensive but decreasing use of incorporating punishment possibilities, predominantly not in the form of peer-punishment, but centralized punishment or redistribution. (2) Rule designers strongly incorporate communicative elements, not in the form of peer-to-peer communication, but rather in appeals to moral sentiments or a framing of the game situation. (3) Rule designers prefer to conceal individual contributions and rather communicate aggregated contributions levels.

[10] The strategy method was applied by Selten, Mitzkewitz and Uhlich (1997) in the framework of a Cournot duopoly, by Keser and Gardner (1999) for a common pool resource game, and by Selten, Abbink, Buchta, and Sadrieh (2003) for a 3x3 game. The famous study by Axelrod (1984) and the study by Keser (2000) are also in this spirit.

[11] Meanwhile the term *strategy method* is commonly used for an experimental protocol in which game playing subjects provide a complete list of actions prior to knowing the actual decision(s) of the other subject(s). It is applied to acquire knowledge about the subject's reaction to *all* possible actions of the other subject(s) and not only to the specifically chosen one(s).

II. Model and experimental design

We recruited 24 students of economics at the University of Erfurt for two separate seminar-type courses (12 in each seminar). The seminars ran over 70 days from April to July 2007. Upon arrival, students were allocated to rule-development groups of four who stayed together for the entire seminar. The course of each seminar is summarized in Table 1.

Time	Playing stages	Rule development
	Preliminary meeting	
Day 1:	play of the basic public goods game	*development groups formed*
Days 2-6:		development of 1st set of rules
Day 7:		handing-in of rules
Days 8-13:	software implementation (experimenter)	
Day 14:	play under 1st rule sets	
Days 15-20:		development of 2nd set of rules
Day 21:		handing-in of rules
Day 22-27:	software implementation (experimenter)	
Day 28:	play under 2nd rule sets	
Days 29-34:		development of 3rd set of rules
Day 35:		handing-in of rules
Days 36-41	software implementation (experimenter)	
Day 42:	play under 3rd rule sets (double weighting)	
Days 43-48:		4th set of rules
Day 49:		handing-in of rules
Days 50-55:	software implementation (experimenter)	
Day 56:	play under 4th rule sets (triple weighting)	
Day 70:	*Final meeting; discussion*	

Table 1: Schedule of the seminars

During a preliminary meeting, potential seminar participants were introduced to the schedule. However, they were not informed on what game they would be going to face during the seminar.[12] The seminar was not accompanied by any lectures on theoretic or experimental investigations into social dilemmas, neither in this term nor in previous terms. On the first day of the seminar (Day 1 in Table 1), the participants experienced the basic game by playing it in a laboratory setting over 25 rounds in a partner design. The basic game was a standard public goods game with four players. Each player *i* contributed an integer amount x_i ($0 \leq x_i \leq 20$) from an endowment of 20 tokens to a common project and kept the remainder. The sum of all contributions was multiplied by 1.6 and divided evenly amongst the players, so that the public good exhibited a constant marginal per-capita

[12] See Appendix A for the information passed to those present at the preliminary meeting, as well as to the instructions for the basic game.

return of 0.4. The resulting payoff function of subject *i* was therefore:

$$\Pi_i = 20 - x_i + 0.4\sum x_j$$

Subjects were informed only about the sum of contributions, not about individual contributions and there was neither punishment nor any other additional non-standard rule feature.

Rule design

At the end of the first meeting, participants were randomly allocated to *design groups* of four students each who stayed together until the end of the seminar. The participants had one week to develop their own set of rules for the public goods game within these design groups. There was no pre-defined "menu" of rules and the subjects were free to develop whatever rules they wanted (in the boundaries of standard rules of ethics). Each rule (component) is potentially attached to costs. The costs reflect the expenditures the implementation of such a rule would imply in a real-world setting. For some rule sets, costs were split into a fixed and a variable part: e.g. setting up the infrastructure to make public announcements as a fixed cost and the variable costs of actually making announcements. The most severe and therefore most expensive interventions are restrictions of the players' action space, e.g. to exclude complete free-riding or to even enforce full contribution. These kinds of coercion would not only require a lot of enforcement power, but severely change the nature of the game and therefore, the cost of full-contribution enforcement was set to the maximum gains to be achieved by such an intervention, i.e. 1200 points. On the other end of the cost range are simple announcements or advertisements to the players.[13] The cost scheme was hidden from the groups to avoid an "anchoring effect", but any subject was given the possibility to ask for the costs of a specific rule set at all times.

Implementation, play, and feedback

At the end of the week, each design-group had to hand in a verbal description of their rule set which was subsequently implemented in the experimental software z-tree[14] and translated into neutrally-worded instructions by the experimenter.[15] After another week, we met again for the first round of play under those rule sets. Randomly allocated *play groups* of 4 played under the different rule sets developed by the rule design groups. To avoid rule designs tailored to a specific subject population, the play groups differed from the design groups. To provide incentives for creating efficient rules and improve them as best as possible, the designers' "payment"[16] was relative to the social benefits

[13] A detailed listing of the introduced rules and the attached costs is provided in Table B1 in appendix B.

[14] Fischbacher (2007).

[15] Subjects were given the possibility to have loaded instructions distributed, incurring the same rule costs as on-screen announcements during the experiment.

[16] As in the previous applications of the strategy seminar we also „paid“ seminar participants in grades correlating to

their institutional rule delivered to the play groups. More precisely, we measured efficiency as the sum of individual payoffs within the play group (who also had to bear the costs caused by the rule-set), as a fraction of the payoffs in the social optimum.

By the end of each round of play, seminar participants were provided with detailed feedback on the performance of all rules within their seminar group by round and group, comprising (i) individual contributions, (ii) efficiency, (iii) returns from the public good abstracting from any costs, and (iv) variable costs, as well as the instructions for all rule sets.[17] Overall, there were five rounds of play, with rounds two to five played under rules developed by the design groups.[18] In order to leave some room for initial experimentation, later rules were weighted higher than earlier rules. Specifically, round two and three are weighted by 1, while round four is weighted by 2 and the final fifth round is weighted by 3. In addition to their rule sets' performance, the profits gained in students' individual play also made up for the seminar mark, so that sabotage of alien rules would be costly to the saboteur. Note that individual performance in the different tournaments was weighted evenly.

We had the students clearly separated into two distinct groups that met at different times, asking them not to communicate with students from the other group on the topic of the seminar. This was done to see whether the rule sets would "converge" to similar sets. Not only did rules not "converge" to the same set over the seminars; they did not even converge within a seminar group.

Data base

In our experiment we obtain a data base of 24 rule sets from two seminar groups composed of 12 participants each. The corresponding three rule groups in each seminar interacted in four tournaments. Out of these 24 rule sets, we excluded two for our analysis, leaving us with 22 data points.[19]

III. Rules used and disregarded rule features

The rules introduced by our subjects are summarized in table B.1 in appendix B. In that table, we briefly describe the rule sets, providing the contribution and efficiency level achieved as well as the costs they gave rise to, grouped by the seminar group and the tournament number. In order to have a better understanding of what the determinants of those rule sets were, we classified the rule components along different categories.

achieved points.

[17] Fixed rule costs were stated in the instructions and implied in the efficiency figures.

[18] Recall that there was no rule performance to be measured in the first round of play, given in this round, subjects played the standard public goods game without any additional rules.

[19] One group redistributed any contributed points surpassing the minimum contribution back to the contributing player before the sum of contributions was multiplied by the public-good factor. This changes the game from being a social dilemma to a coordination game. We excluded this rule set, which was applied by the group in tournament 2 and 3.

1. Punishment and redistribution

The role of punishment in social dilemma situations is widely and prominently discussed in the literature[20]. Although our subjects were not trained in this literature, many of the submitted rule sets made use of punishment or redistribution features. Interestingly however, only two of them employed a peer-to-peer mechanism. Most often, rules specified the deduction of points from the lowest-contributor and, in case of redistributive mechanisms, the reallocation of these points in favor of the highest-contributing player. In some of the cases, deducted points were transferred to an account that was to be redistributed at the end to the player having contributed the most. One rule set incorporated a 'warning system': those contributing less than the mean were asked to increase their contributions in the following period. In case this advice was not followed, they were punished by an amount conditional on the earlier deviation from the mean. Most of the sets, however, had a direct, automatic punishment or redistribution mechanism based on contribution ranks, with differentiated punishment of the lower-contributors. These observations give rise to four (not mutually exclusive) components capturing different punishment and redistribution techniques. The components are formulated in such a way that the question of whether a rule set satisfies it can unambiguously be answered by yes or no.

> *1a. Punishment and redistribution (pun)*: the rule set provides for either destruction of a part of a player *i*'s points (punishment), or a reallocation thereof in favor of at least one other player *j* (redistribution), conditional on *i*'s behavior.

This condition was fulfilled for 75% of all rule sets. Remarkably, in the first tournament, all rule sets include either punishment or redistribution, or both. However, only half the sets feature any of them in the final tournament, and those that do, use redistribution. In other words, 'pure' punishment is no longer observed in the final tournament, but "punishment" through redistribution or rule cost assignment conditional on the player's contribution is. While this could be attributed to details in the cost scheme implemented, it cannot be ignored that the frequency of use of these mechanisms steadily declines over time.

To obtain a better understanding of the *pun* rules, we introduce three further classifying variables, the first of which relates to the frequency with which the deduction regimes played a role in the game:

> *1b. Punishment or redistribution in every period (punEP)*: punishment or redistribution (as described in 1a.) takes place in every single round.

[20] See e.g. Yamagishi (1986), Ostrom, Walker, Gardner (1992), Fehr and Gächter (2000, 2002), Denant-Boemont et al. (2007), Nikiforakis (2008), Carpenter and Matthews (2009) for experimental studies, and Henrich and Boyd (2001), Boyd, Gintis, Bowles, and Richerson (2003), Hauert et al. (2007), Dreber et al. (2008) for theoretical approaches.

This category makes the distinction between rules with roundly punishment or redistribution and those implementing them only periodically or even only once. A rule set exhibiting *punEP* provides for roundly deduction of points, a characteristic that was displayed by 13 out of 24 rule sets. In terms of rule sets which exhibit characteristic *pun*, the fraction of rule sets with *punEP* decreases from six out of six (eight out of ten in tournaments one and two) to two out of three (five out of eight in tournaments three and four).

Another distinction can be made with regard to players' influence on the points to be deducted:

1c. Redistribution endogeneity (redEnd): punishment or redistribution (as described in 1a.) is administered by players themselves rather than automatically.

Most rule sets (16 out of 18 rule sets with *pun*) proposed by our rule groups deprived the players from any influence on the points to be deducted, once the contribution decisions had been taken. Only in one instance, punishment was in the form of peer-to-peer punishment as in the typical public goods experiments with punishment,[21] and in the other rule set, players contributing more than 14 tokens were allowed to jointly decide (through a voting procedure) on the allocation of rule costs among the remaining players.

Finally, we distinguish rules that involve a single, concentrated reward payment at the end of a session as a special case of a redistributive rule:[22]

1d. 'Big bonus' (BB): redistribution takes on the form of a fund paid into and one distributive action (the allocation of a 'jackpot') at the end of round 25.

There were three rule sets that involved roundly payments into a fund that was awarded to the highest-contributing player at the end of the respective session (with an equal-split rule in case of a tie). All of them were designed by the same rule group.

2. Feedback on individual behavior

Several studies (e.g., Milinski, et al., 2006, Sommerfeld, et al., 2007) have shown that reputation-building opportunities may be very effective in promoting cooperation. A necessary prerequisite for being able to build a reputation is having some kind of identity. Bohnet and Frey (1999) show that identification alone suffices to increase the degree of pro-social choices in prisoner's dilemma and dictator games. Rule sets could allow for identification, e.g. in terms of a fixed player ID to subjects, but in our seminars, this happened only in a minority of the cases.

[21] E.g., Fehr and Gächter (2000, 2002).

[22] Recall that rule sets involving a 'big bonus', like all other rule sets, could not be designed in a way that would make the defecting equilibrium disappear.

2a. Feedback on individual behavior (ID): players are informed not only on the sum of contributions, but also on individual contributions.

In fact, subjects were provided with identity-based feedback only in seven rule sets. One set provided detailed feedback on contributions without identity numbers, and the remaining two thirds of the rule sets did not inform players on individual contributions. In two cases, players faced a rather special situation we also want to separate: players under these two rule-sets did not receive any feedback on contributions at all.

2b. No feedback on contributions (noFB): the sum of contributions is concealed from the players.

In both cases, this rule was coupled with a 'big malus', a severe punishment action of the lowest-overall-contributor at the end of the game, unless a pre-specified cumulative contribution level had been achieved by the group.[23] In the first instance of the rule-set, a considerable amount of contributions was achieved (1271 tokens, as compared to the 834 required to circumvent the 'big malus'). Subsequently, the group modified their rule-set, asking for a group-contribution of 1667 tokens, which the play-group failed to comply with by a margin of 120. The application of the 'big malus' led to a disastrous result in terms of efficiency and the group abandoned this strategy.

3. Communication: subject-to-subject and "lawmaker"-to-“people” (framing)

Another tool that has proven to be very effective in inducing cooperation in social-dilemma situations is communication (e.g. Isaac and Walker, 1988, Ostrom, Gardner and Walker, 1994, Cason and Khan, 1999, or Brosig, Weimann and Ockenfels, 2003). Indeed, our subjects also implemented forms of communication, but on a vertical rather than a horizontal level. This means that they mostly abstained from peer-to-peer-communication and instead used a communication of the lawmaker to the players.

3a. Player-to-player communication (CommP2P): there is some form of communication between players, e.g. in form of a signaling opportunity such as in cheap-talk agreements.

No group implemented an open-communication component in the form of a virtual chat-room. Only in the final tournament, one rule group in each seminar opted to allow players to engage in signaling behavior through a unanimity vote on a (non-binding) covenant, with roundly renewal (contingent on that there have not been more than two breaches in the past) and requiring subjects to contribute fully in one case, and periodical votes coupled with a proposed minimum of 15 tokens in the

[23] We do not introduce an additional 'big malus' category, as the two rule-sets are already uniquely classified by the *noFB* category. No other rule-set employed a 'big malus', and thus, such a category would not add any information.

other.[24]

On the other hand, communication by the lawmakers to the players was common. Communication of this form was implemented in terms of (moral) appeals or a framing of the public goods situation.[25] This is especially remarkable because all subjects were – by then – experienced players of laboratory public goods provision and this was common knowledge. Furthermore, when playing under those rules they knew that the framing was artificially imposed and had no relevance to the game actually played.

3b. Some frame (frame): the game is played after the issuing of an appeal (to moral sentiments, the advantageousness of the social optimum, or general social-welfare considerations), with a background story, or even with feedback conditional on behavior.

Less than half of all rule sets did without a framing (13 out of 24 had some framing). Rule components classified as falling into this category include priming attempts, such as the display of roundly changing statements similar to "One cannot live without trusting others". Further, they include individual feedback conditional on behavior such as messages "To live means believing in something. In our case, in the community, thanks for that!" in case of a full contribution. Or framing in the narrower sense, embedding the contribution decision in contexts like the building of a school in Afghanistan or public goods arising in a local neighborhood.

To further distinguish rules making use of a *frame* component, we separate those containing a one-time appeal at the beginning only from those envisaging repeated messages.

3c. Frame in every period (frameEP): an appeal, a background story, or behavior-conditional feedback is displayed in every single round.

Out of the 13 rule-sets containing a *frame*, six also had *frameEP* as a characteristic.

4. Participative elements

Experimental studies like Ostrom, Walker and Gardner (1992) and Tyran and Feld (2006) have shown that increased participation opportunities, such that players may influence the situation they are facing, lead to more cooperative outcomes. We were curious to find out whether our subjects would anticipate this and implement features of endogenous rule adaption or even rule change.

4. Endogeneity (end): the rule-set provides for institutional change as a consequence of either player behavior or a vote.

[24] The latter rule group envisaged the possibility of changing the minimum requirement in case of general adherence to the covenant; however, their covenant did not succeed in inducing the expected cooperation.

[25] See Ross and Ward (1996) and Cookson (2000) on positive effects of framing, for studies not finding a framing effect, see Rutte, Wilke and Messick (1987) or Rege and Telle (2004).

Indeed, such features are used increasingly towards the end. However, only one fifth of all rule sets contain them and in one of the seminar groups, such elements are introduced only in the last tournament. Examples of endogenous rule features are a vote on removing any features that go beyond the basic game after the eighth round or the introduction of peer punishment contingent on players paying for a monitoring institution.

5. Omitted rule elements

The list of rule components that were discussed within the rule design phase but not implemented contains quite invasive rules. Groups considered, for example, to enforce a certain contribution on subjects or to completely exclude non-cooperators from the game. Such harsh rules, however, never made it from the group discussion to the actual implementation, although they were not prohibitively expensive.

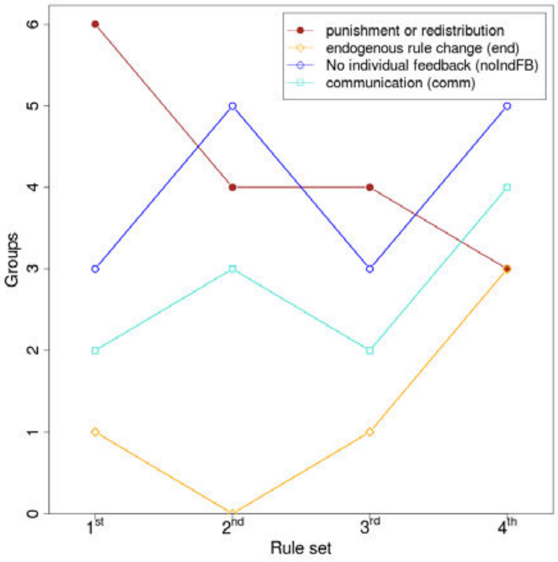

Figure 1: Frequency of components by rule set.

The incidence of the rule components summarized under the four component categories traced over the course of the seminar is depicted in Figure 1. As mentioned before, we can see a clear trend away from *pun* rules, as well as from individual feedback, which was not a widespread rule component even in the beginning. Communication in the form of frames or appeals as well as of signaling opportunities, on the other hand, was used increasingly, as was endogenous rule change. [26]

In a second step, we looked at typical combinations of rule components – or at combinations that

[26] In Figure 1, as in the ensuing analysis, we re-aggregate the subcomponents to main components, as this is enough to convey the important insights. The same steps of analysis using the subcomponents were taken for an earlier working-paper version of this paper and can be found in appendix C. For the feedback component we use "no feedback on individual behavior (*noIndFB*)" as the main category, while for the communication component, we define every rule set to fall into the main category "communication (*comm*)" that has either *commP2P* or *frame*.

typically would not be combined by our subjects.[27] The main result to be taken from this analysis is that all combinations but one are implemented more than once; the only combination that has not been implemented was that of communication (*comm*) coupled with an individual-feedback (*ID*). In other words, those engaging in the psychological techniques of framing, priming, and appealing explicitly choose to render others' individual behavior opaque.

A table classifying the different rule sets according to our characteristics can be found in Appendix B (see Table B2). The classification is also used for the typicity analysis reported in section V.[28] Before we do that analysis, we explore the performance of rules exhibiting a certain component in section IV, in order to assess whether the rule-adjustment process went 'in the right direction'.

IV. Rule performance and rule adaptation

In the preceding section, we categorized subjects' rule sets and established the frequencies with which rule components were used in the different tournaments. In this section, we set out to evaluate those components' contribution to the performance of the rule set they are incorporated in.

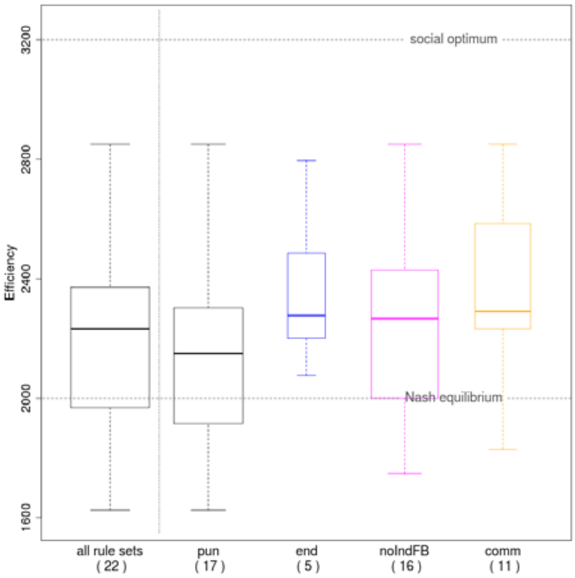

Figure 2: Boxplots of achieved efficiency levels, by rule components

Figure 2 gives a first hint at an answer, displaying boxplots of the efficiency levels achieved by all rule-sets employing a certain main component. Judging by Figure 2, one may argue that having characteristics like endogeneity (*end*), no individual feedback (*noIndFB*) or communication (*comm*) in one's rule set seems to be a good idea, while employing punishment (*pun*) does not seem to lead

[27] Tables summarizing this analysis can be found in appendix C.

[28] To run a typicity analysis, characteristics displayed by less than half of all strategies have to be reverted (such that, e.g., *ID* becomes *noID*, i.e., "subjects are *not* informed on individual contributions but only on their sum"), for technical reasons outlined in Kuon (1993).

to a good performance. To examine this conjecture more closely, we analyze our rule-sets using a simple distribution-free measure of success. We rank the rule sets by efficiency (with the highest rank corresponding to the highest efficiency) and calculate the average rank of all rule sets displaying a certain characteristic. Comparing this to the median rank of all rule sets provides us with an index for the discriminatory power of the component. Division by *(n-1)/2*, where *n* is the total number of rule sets, finally normalizes the index to lie within the interval [-1,1]. This index measures a component's contribution to the efficiency gained by the rule set. This is why we term the index shortly as the *component-efficiency index*. A component-efficiency index of 1 implies that only the best-performing rule set displays this component, while a component-efficiency index of -1 means that the characteristic is only made use of by the worst-performing rule set. A characteristic that is employed in every strategy would lead to a component-efficiency index of 0, as would any component for which the sum of rank deviations from the median above the median is equal to the sum of those below it. If R_i denotes the rank of rule set *i*, and Y_j is the set of all rule sets that exhibit characteristic *j*, the component-efficiency index ρ_j can be expressed as follows:

$$\rho_j = \begin{cases} \frac{2}{n-1}(\sum_{y \in Y_j} \frac{R_y}{|Y_j|} - \frac{n+1}{2}), & |Y_j| > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This formulation of a component-efficiency index is a rather simple approach assuming a linear separability of the performance of rule sets, neglecting any interaction effects between rule components. Note further that the index is based on a relative ranking of marginal component effects, and can therefore only explain deviations from the median efficiency rank. Nevertheless, a virtual rank ordering of rule sets by simply adding up the components' ρ_j yields a surprisingly good prediction for the rank ordering by efficiency: the Spearman correlation coefficient between the two rankings is $r_s = 0.49$ ($p = 0.0214$). If we take into account that not all components contribute to efficiency to the same extend, this is a surprisingly high correlation.[29] In Figure 3, we depict the component-efficiency indices for our rule characteristics, grouped by tournaments. The two components that seem to boast efficiency the most seem to be the use of communication (*comm*) and giving players a way to influence the rules they are playing under (*end*), while *not* providing subjects with feedback on individual contributions (*noIndFB*) seems to enhance efficiency only slightly. In contrast, *pun* features do not seem to make up a clear picture across tournaments: while they tend to be more prevalent among worse-performing rule sets, this is notably different in the

[29] Running a mixed-effects estimate of normalised rank deviations from the median rank on the indices (with random effects for rounds, the interaction of rounds and seminar group, and rule groups), the rank correlation between the ensuing fitted ranks and the true ranks can be driven up further ($r_s = 0.88$, $p < 0.001$).

final tournament. It seems that our subjects have learnt to design efficient *pun* mechanisms by the end of the experiment. On the other hand, we see that the seemingly successful mechanisms of *comm* and *end* (Figure 3.c) do not perform as well compared to final-set *pun* rules (Figure 3.b). When compared to rule-sets from other rounds, however, they still score above average.[30]

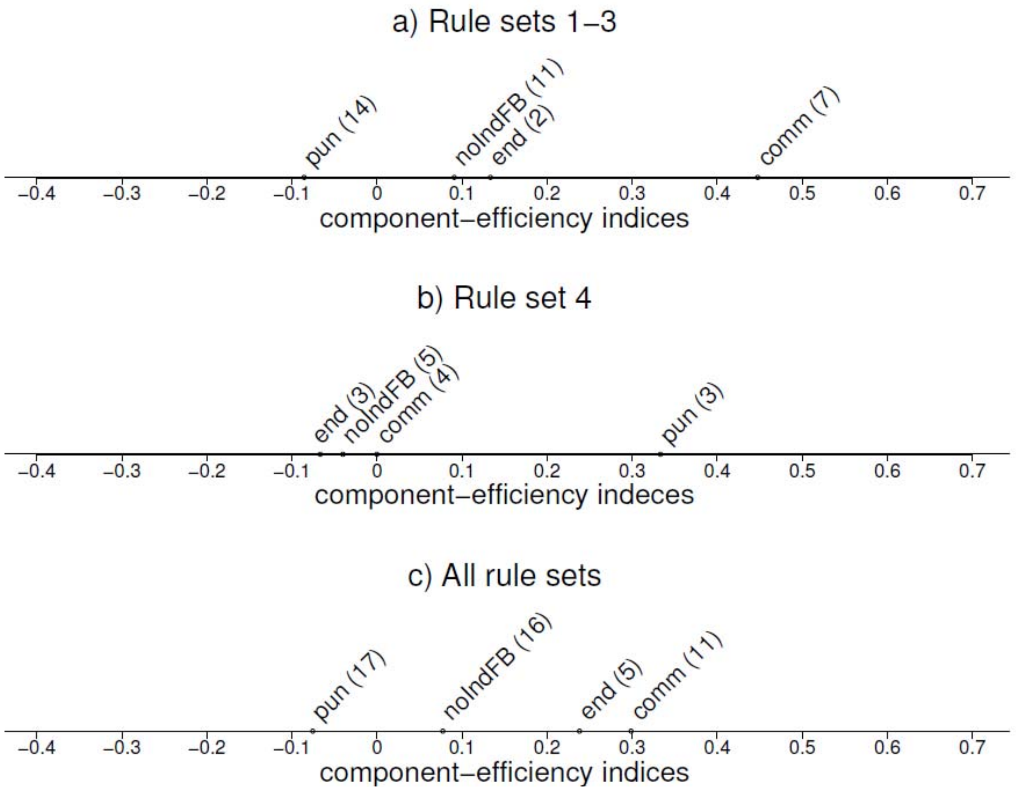

Figure 3: Component-efficiency indices: This index measures a component's contribution to the efficiency gained by the rule set. Panel a) component-efficiency indices for rule sets 1-3. Panel b) component-efficiency indices for rule set 4. Panel c) component-efficiency indices for all rule sets. The numbers of individual rule-sets having a component are given in parentheses; if none, the index is omitted.

Based on our indices as a – however imperfect – measure of success, let us take a look at the rule changes over the tournaments as depicted in Figure 1. We observe that (i) punishment and redistribution systems lead to low efficiency (and only slightly higher contributions), which corresponds with the trend away from such mechanisms. The fact that the decline in the use of *pun* is slow and only partial is surprising especially in seminar I, where there was one rule group that managed to achieve the highest degree of efficiency within their seminar in three out of four tournaments - using rules that did not contain either punishment or redistribution. In fact, this was the only rule group to renounce the use of the *pun* component within this seminar group. In other

[30] Contrasting the component-efficiency indices to indices calculated in the same way but taking contribution levels instead of efficiency levels as the underlying metric, we find very similar results. We take this an indication for the robustness of our rule-cost scheme.

words, participants in the remaining rule groups fail to recognize that not making use of punishment may actually be a competitive advantage: even in view of the superior results achieved by their competitor group, they fail to mirror that group's superior strategy.[31] In seminar II, rule sets abstaining from a ’monetary’ sanctioning of behavior do not perform as well; in two out of three instances, they achieve only the second highest efficiency, the third set following one of the two just mentioned on third place.

In the end, is the use of punishment a bad idea? Maybe not: *pun* mechanisms as designed in the last round (i.e., redistribution of rule costs as the only such mechanism) increase efficiency.[32] From Figure 1, we also know that the use of ID-based feedback tends to decline over time. Figure 3 shows that this is not as surprising as it may seem: the efficiency index for the first three rounds is $\rho_{noIndFB,1-3} = 0.09$. Hence, ID-based feedback does not seem to have a positive effect on cooperation as long as there is no explicit award or commendation, or the possibility to be rewarded in an ancillary game as e.g. in Milinski, et al. (2006). Correspondingly, we observe the stated trend away from ID-based feedback, with three rule sets having the *noIndFB* component in the first and five sets in the final tournament.

Turning to components with some communicative element, be it between subjects or between the 'lawmakers' and their 'people', we observe a steady trend towards *comm* components. This corresponds well with the positive effect such components have in the first three rounds ($\rho_{comm,1-3} = 0.45$). Finally, having the 'people' participate in the shaping of their world through endogenous features tends to foster efficiency ($\rho_{end,1-3} = 0.24$).

Do 'lawmakers' change their rules for the better?

The final tournament was the most decisive for the students’ success in the seminar and the intermediate tournaments were meant to improve the rule sets and their lower weights in the final grading enable experimentation. Participants have successfully used this opportunity. The median efficiency from the third to the fourth (final) tournament increases significantly ($p_{34} = 0.0625$, pair-wise two-tailed Wilcoxon signed-ranks test, where p_{xy} denotes the p-value for the test between tournament x and y), whereas it does not between the first and the second and between the second and the third, respectively ($p_{12} = 0.3125$ and $p_{23} = 0.625$). What did the rule groups do to achieve

[31] This reluctance to follow the good example of their competitor group could, of course, easily be explained by postulating a psychological bias to focus on contributions. However, from our discussions with the subjects we got the impression that they were well-aware of the fact that high(er) contribution averages do not necessarily lead to higher efficiency, but have to be measured against potentially increasing rule costs.

[32] Unfortunately, a meaningful statistical comparison using e.g. conditional average payoffs is not possible, given rules containing a feature and those not containing this feature within a seminar group are not independent.

this improvement? To find an answer to this question and possibly shed some light on typical mistakes and typical improvements, we performed a typicity analysis based on the characteristics introduced in section III.

V. Typicity analysis

The preceding analyses considered the rule components in isolation and estimated their contribution to the rules' efficiency. In the following we analyze the interdependencies of the rule sets' characteristics with a typicity analysis. This type of analysis, introduced by Kuon (1993) and successfully implemented e.g. by Selten, Mitzkewitz and Uhlich (1997), is used to determine what a 'typical strategy' or, in our case, a 'typical rule set' is and what the typical characteristics of these rule sets are. The two measures are interdependently intertwined: the typical rule sets are the ones that carry the typical characteristics and the typical characteristics are the ones that occur in typical strategies. Technically the typicity analysis is the solution of an Eigen value problem, providing weights for the rule sets and the characteristics. Typicities of characteristics add up to one and are all equal to one divided by the number of characteristics if the number of strategies containing each of these features is the same, such that the reciprocal value of the number of characteristics forms the natural reference point against which typicities may be evaluated. On the other hand, the more typical a rule set's features are, the more typical is the rule set itself. More precisely, the typicity of a rule set is the sum of its characteristics' typicities.[33] Once the typicities of rule sets and characteristics are established, we can look for whether typicity is in any way correlated with efficiency and thus assess how well-targeted our subjects were in terms of their search for better solutions. Given we have found rule sets in the final tournament to perform better than those in the preceding tournaments, we run two distinct analyses for these two sets and compare their typicities to find out how the typical rule sets differed. Table 2 lists the corresponding component typicities.

	pun	noIndFB	comm	noEnd
$\rho_{j,1-3}$	-0,086	0,091	0,448	-0,019
$\rho_{j,4}$	0,333	-0,040	0,000	0,067
Rule sets 1-3	0,277	0,255	0,168	0,300
Rule set 4	0,176	0,333	0,285	0,206

Table 2: Performance indices, as introduced in section IV (lines 1 and 2); rule component typicities over the first three and the final rule sets (lines 3 and 4). Note that the *end* characteristic has been reverted for the purpose of the typicity analysis, because – for technical reasons – the characteristics have to be formulated such that the majority of rule sets carry it. The performance indices reported

[33] For a thorough discussion of the mathematical properties of this method, cf. Kuon (1993).

in lines 1 and 2 correspond to the changed component formulation and may hence be different from those reported in Section IV.

What we see in Table 2 is that the change in rule component typicities largely corresponds to the frequency changes we know from Figure 1. Also from Table 2, we observe that a positive component performance in rule sets 1 to 3 (i.e., $\rho_{j,1-3} > 0$) tends to lead to an increase in the component's typicity. Furthermore, we find an increasing diversity in rule sets: average rule-set typicity declines from 0.888 in the first rule-set round to 0.773 in the final one even if we base these averages upon final-rule-set typicities. This is a clear sign of a 'divergence' of rule sets.

A correlation analysis reveals that there is no systematic relation between rule typicity on the one hand and either contributions or efficiency, on the other. This lack of a systematic correlation between rule typicity and efficiency may be due to the presence of initial "typical mistakes" that are only hesitatingly done away with. "Typical improvements" fall into two categories: on the one hand, *pun* rules are improved, decreasing their wastefulness through a switch from pure punishment to redistribution rules; on the other, we observe the increasing typicity of communication and participative features, as well as the decline in the typicity of ID-based feedback.

VI. Discussion and Implications

This paper reports an experiment in which subjects act as rule makers for a social dilemma game. The novelty of the approach lies in endogenizing the institutional design process by transferring it to subjects who are free in "inventing" any institutional setting and not bound to pre-specified rules. The designers' incentive for acting as "good" rule makers is that their payment strongly depends on the social benefits that accrue from play under the designed rule. This experimental approach allows studying the rules that social planners facing a social dilemma "invent" and the development of these rules over time. It expands our view on rules fostering cooperation and guides to important new avenues for future research.

We can summarize our findings as follows: (1) Institution designers combine two or more rule components instead of relying on single-component rule sets. (2) Designers regularly include appeals to moral sentiments or a framing of the game situation. (3) Designers predominantly choose rules that render information on individual contributions opaque but often provide aggregate contribution information instead. (4) Designers make extensive but decreasing use of punishment possibilities, predominantly not in the form of peer-punishment, but in the form of centralized punishment or redistribution. (5) Designers do not incorporate the possibility of assigning leaders, or ostracism.

The finding that institution designers frequently and successfully use framing is particularly noteworthy. The insight that the framing of the game may influence subjects' behavior is not new. Examples are Sonnemans, Schram, and Offerman (1998) who compare a 'public goods' game with an equivalent 'public bad' game, or Andreoni (1995) comparing a positive-frame investment in the public good and a negative-frame purchase of the private good, or framing the identical prisoner's dilemma games as either the "Wall Street Game" or the "Community Game" (Liberman, Samuels, and Ross 2004). However, in our experiment, the subjects playing under these rule sets were experienced players, equipped with a deep understanding of the logic of the game and completely aware that the chosen frames have no connection to reality. Nonetheless, framing was increasingly chosen and a successful means in achieving efficiency. While one of the groups opted to tell its subjects the public good was a school in Afghanistan, another group had its subjects play in a virtual neighborhood. Yet other rule sets displayed a "slogan of the round" such as: "Only within the community, people are beings conscious of their strength", or moral appeals directly asking subjects to contribute. The success of these design components shows that attempts to activate social norms through appeals or even general moral statements tend to be more than "just words": they seem to foster cooperation even amongst casehardened addressees. On a more general level this calls for further research to better understand the advantages and downsides of frames and appeals. A recent example is the study by Cohn, Fehr, and Maréchal (2014), who show that a significant proportion of bank employees become dishonest when their identity as bankers is rendered salient.

A second noteworthy finding is that designers render information on the other players' provisions rather opaque - and that this tends to be a successful strategy. Conditionally cooperative subjects align their provision with what they believe others will contribute. Providing them with detailed information on past contributions of their peers may yield the most precise basis for the calculations. Yet, institution designers increasingly preferred to establish a certain degree of opaqueness by just providing the average contribution, as this turned out to be successful in enhancing cooperation. This is well in line with the observations of Hoffmann, Lauer, and Rockenbach (2013), which show that not providing the above average contributors with feedback about the low contributions of others stabilized contributions to a public good on a high level. Nonetheless, hiding detailed contribution information and traceable identifications excludes reputation concerns that have been shown to be powerful in increasing cooperation (Milinski et al. 2002). The interaction of these effects is not yet fully understood and calls for future research.

A third noteworthy observation is that punishment, in various disguises, was a focal design element in almost all initial rule sets. This finding underlines the importance of punishment rules in public goods settings. Remarkably, however, punishment mechanisms were not designed in the form of peer punishment, but rather in the form of pre-specified rules of deduction and/or redistribution contingent on complying with provision targets. These provision targets were either fixed levels (e.g. full provision) or contingent on the other group members (e.g. not being the lowest-contributing player). This third observation parallels a scholarly interest in centralized punishment mechanisms that has gained momentum only very recently (with the two main lines of research sparked off by Putterman, Tyran, and Kamei, 2011, and Sigmund, De Silva, Traulsen, and Hauert, 2010). This research suggests that whether centralized punishment is preferred over peer-punishment may be influenced strongly by the details of the punishment schemes (e.g., whether semi-centralized "pool punishment" includes second-order punishment of non-punishing contributors, cf. Zhang et al., 2014), and boosts the idea that its effectiveness is mediated strongly by whether it is implemented by a vote (e.g., Markussen, Putterman, and Tyran, 2012). Yet other important determinants of a preference for centralized punishment that may play out in our design are the higher degree of control the social planner has over the 'executive authority', or a belief that clearly specified rules will foster cooperation better than (arbitrary) peer punishment. To determine the relative importance of each of these factors is an important question to be addressed by future research.

References

Andreoni, J. (1995). "Warm-Glow Versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments." Quarterly Journal of Economics **110**(1): 1-21.

Arbak, E. and Villeval, M.-C. (2013). "Voluntary Leadership: Motivation and Influence." Social Choice and Welfare **40**(3): 635-662.

Axelrod, R. (1984). The Evolution of Cooperation. New York: Basic Books.

Bochet, O., Page, T. and Putterman, L. (2006). "Communication and Punishment in Voluntary Contribution Experiments." Journal of Economic Behavior & Organization **60**(1):11-26.

Bohnet, I. and Frey, B. (1999). "The Sound of Silence in Prisoner's Dilemma and Dictator Games." Journal of Economic Behavior and Organization **38(1)**: 43-57.

Boyd, R., Gintis, H., Bowles, S. and Richerson, P. (2003). "The Evolution of Altruistic Punishment." Proceedings of the National Academy of Sciences of the United States of America **100**(6): 3531-3535.

Brosig, J., Weimann, J. and Ockenfels, A. (2003). "The Effect of Communication Media on Cooperation." German Economic Review **4**(2): 217-243.

Carpenter, J. and Matthews, P. (2009). "What Norms Trigger Punishment?" Experimental Economics **12**(3): 272--288.

Cason, T. N. and Khan, F. U. (1999). "A Laboratory Study of Voluntary Public Goods Provision with Imperfect Monitoring and Communication." Journal of Development Economics **58**(2): 533-552.

Cinyabuguma, M., Page, T. and Putterman, L. (2005). "Cooperation under the Threat of Expulsion in a Public Goods Experiment." Journal of Public Economics **89**(8): 1421-1435.

Cohn, A., Fehr, E. and Maréchal, M. A. (2014). "Business Culture and Dishonesty in the Banking Industry." Nature **516:** 86-89.

Cookson, R. (2000). "Framing Effects in Public Goods Experiments." Experimental Economics **3**(1): 55-79.

Davis, D. D. and Holt, C. A. (1993). Experimental Economics. Princeton: Princeton University Press.

Dreber, A., Rand, D. G., Fudenberg, D. and Nowak, M. A. (2008). "Winners Don't Punish." Nature **452**(7185): 348-351.

Denant-Boemont, L., Masclet, D. and Noussair, C. (2007). "Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment." Economic Theory **33**(1): 145-167.

Falk, A., Fehr, E. and Fischbacher, U. (2005). "Driving Forces behind Informal Sanctions." Econometrica, **73**(6): 2017-2030

Fehr, E. and Gächter, S. (2000). "Cooperation and Punishment in Public Goods Experiments." American Economic Review **90(4)**: 980-994.

Fehr, E. and Gächter, S. (2002). "Altruistic Punishment in Humans." Nature **415**: 137-150.

Fischbacher, U. (2007). "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." Experimental Economics **10**(2): 171-178.

Fischbacher, U. and Gächter, S. (2010). "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments." American Economic Review **100**(1): 541-56.

Guillen, P., Schwieren, C. and Staffiero, G. (2007). "Why Feed the Leviathan?" Public Choice **130(1-2)**: 115-128.

Gürerk, Ö., Irlenbusch, B. and Rockenbach, B. (2006). "The Competitive Advantage of Sanctioning Institutions." Science **312**: 108-111.

Gürerk, Ö., Irlenbusch, B. and Rockenbach, B. (2014). "On Cooperation in Open Communities." Journal of Public Economic **120**: 220-230.

Güth, W., Levati, M. V., Sutter, M. and van der Heijden, E. (2007). "Leading by Example with and without Exclusion Power in Voluntary Contribution Experiments." Journal of Public Economics **91**(5-6): 1023-1042.

Hardin, G. (1968). "The Tragedy of the Commons." Science **162**(3859): 1243-1248.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A. and Sigmund, K. (2007). "Via Freedom to Coercion: The Emergence of Costly Punishment." Science **316**(5833): 1905-1907.

Henrich, J. and Boyd, R. (2001). "Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas." Journal of Theoretical Biology **208**: 79-89.

Hoffmann, M., Lauer, T. and Rockenbach, B. (2013). "The Royal Lie." Journal of Economic Behavior and Organization **93:** 305-313.

Isaac, M. R. and Walker, J. M. (1988). "Communication and Free-Riding Behavior: The Voluntary Contributions Mechanism." Economic Inquiry **26**: 585-608.

Keser, C. (2000). "Strategically Planned Behavior in Public Good Experiments." CIRANO Scientific Series 2000s-35.

Keser, C. and Gardner, R. (1999). "Strategic Behavior of Experienced Subjects in a Common Pool Resource Game." International Journal of Game Theory **28**(2): 241-252.

Kosfeld, M., Okada, A. and Riedl, A. (2009). "Institution Formation in Public Goods Games." American Economic Review **99**(4): 1335-1355.

Kuon, B. (1993). "Measuring the Typicalness of Behavior." Mathematical Social Sciences **26**(1): 35--49.

Ledyard, J. O. (1995). Public Goods: A Survey of Experimental Research. Handbook of Experimental Economics. Kagel, J. and Roth, A. (Princeton University Press)**:** 111-194.

Liberman, V., Samuels, S. and Ross, L. (2004): "The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves." Personality and Social Psychology Bulletin **30**(9): 1175-1185.

Maier-Rigaud, F. P., Martinsson, P. and Staffiero, G. (2005). "Ostracism and the Provision of a Public Good – Experimental Evidence." Journal of Economic Behavior and Organization **73**(3): 387-395.

Markussen, T., Putterman, L., Tyran, J.-R. (2014). "Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes." Review of Economic Studies **81**(1): 301-324.

Masclet, D., Noussair, C., Tucker, S., and Villeval, M.-C. (2003). "Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism." American Economic Review **93**(1): 366-380.

Milinski M., Semmann D. and Krambeck H.-J. (2002). "Reputation helps solve the 'tragedy of the commons.'" Nature **415**(6870): 424-426.

Milinski, M., Semmann, D., Krambeck, H.-J. and Marotzke, J. (2006). "Stabilizing the Earth's Climate Is Not a Losing Game: Supporting Evidence from Public Goods Experiments." Proceedings of the National Academy of Sciences **103**(11): 3994-3998.

Nikiforakis, N. (2008). "Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" Journal of Public Economics **92**(1-2): 91-112.

Nikiforakis, N. and Normann, H.-T. (2008). "A Comparative Statics Analysis of Punishment in Public-Good Experiments." Experimental Economics **11**(4): 358-369.

Ostrom, E. (1998). "A Behavioral Approach to the Rational Choice Theory of Collective Action: Presidential Address, American Political Science Association, 1997." The American Political Science Review **92**(1): 1-22.

Ostrom, E. (2000). "Collective Action and the Evolution of Social Norms." Journal of Economic Perspectives **14(3)**: 137-158.

Ostrom, E., Gardner, R. and Walker, J. A. (1994). Rules, Games, and Common-Pool Resources. Ann Arbor: The University of Michigan Press.

Ostrom, E., Walker, J. M. and Gardner, R. (1992). "Covenants with and without a Sword: Self-Governance Is Possible." The American Political Science Review **86**(2): 404-417.

Potters, J., Sefton, M. and Vesterlund, L. (2005). "After You--Endogenous Sequencing in Voluntary Contribution Games." Journal of Public Economics **89**(8): 1399-1419.

Putterman, L., Tyran, J.-R. and Kamei, K. (2011). "Public goods and voting on formal sanction schemes." Journal of Public Economics **95**(9-10): 1213-1222.

Rege, M. and Telle, K. (2004). "The Impact of Social Approval and Framing on Cooperation in Public Good Situations." Journal of Public Economics **88**: 1625-1644.

Rockenbach, B. and Milinski, M. (2006). "The Efficient Interaction of Indirect Reciprocity and Costly Punishment." Nature **444**(7120): 718-723.

Ross, L. and Ward, A. (1996). Naïve Realism in Everyday Life: Implications for Social Conflict and

Misunderstanding. Values and Knowledge. Reed, E. S., E.Turiel and Brown, T., Lawrence Erlbaum (Mahwah, NJ)**:** 103-135.

Rutte, C. G., Wilke, H. A. M. and Messick, D. M. (1987). "The Effects of Framing Social Dilemmas as Give-Some or Take-Some Games." British Journal of Social Psychology **26**: 103-108.

Selten, R. (1967). "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments." Beiträge zur experimentellen Wirtschaftsforschung. Sauermann, H. (Tübingen: J.C.B. Mohr (Paul Siebeck)): 136-168.

Selten, R., Abbink, K., Buchta, J. and Sadrieh, A. (2003). "How to Play 3x3 Games – A Strategy Method Experiment." Games and Economic Behavior **45**(1): 19-37.

Selten, R., Mitzkewitz, M. and Uhlich, G. R. (1997). "Duopoly Strategies Programmed by Experienced Players." Econometrica **65**(3): 517-555.

Sigmund, K., De Silva, H., Traulsen, A. and Hauert, C. (2010). “Social Learning Promotes Institutions for Growing the Commons.” Nature **466**: 861-863.

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D. and Milinski, M. (2007). "Gossip as an Alternative for Direct Observation in Games of Indirect Reciprocity." Proceedings of the National Academy of Sciences **104**(44): 17435-17440.

Sonnemans, J., Schram, A. and Offerman, T. (1998). "Public Good Provision and Public Bad Prevention: The Effect of Framing." Journal of Economic Behavior & Organization **34**(1): 143-161.

Sutter, M., Haigner, S. and Kocher, M. G. (2005). "Choosing the Stick or the Carrot? Endogenous Institutional Choice in Social Dilemma Situations." Review of Economic Studies **77**(4): 1540-1566.

Tyran, J.-R. and Feld, L. P. (2006). "Achieving Compliance When Legal Sanctions Are Non-Deterrent." Scandinavian Journal of Economics **108**(1): 135-156.

Vesterlund, L. (2003). "The Informational Value of Sequential Fundraising." Journal of Public Economics **87**(3-4): 627-657

Yamagishi, T. (1986). "The Provision of a Sanctioning System as a Public Good." Journal of Personality and Social Psychology Review **51**(1): 110-116.

Zhang, B., Li, C., De Silva, H., Bednarik, P. and Sigmund, K. (2014). “The Evolution of Sanctioning Institutions: An Experimental Approach to the Social Contract.” Experimental Economics **17**(2): 285-303.

Designing Institutions for Social Dilemmas

Bettina Rockenbach and Irenaeus Wolff

Supporting Information

Appendices

A1. Information on the course of the seminar

Welcome to our simulation-game seminar "social dilemma"!

During this semester, you will have the opportunity to participate in a business-game type of seminar. First, you will gain experience in interacting with other seminar participants in the particular "game situation". Afterwards, you will develop "sets of rules" for the game situation within groups, gain experience with implementing these rules, and have the opportunity to refine these rules.

Schedule:

1st appointment, 19th april, preliminary meeting:

> Today we will clarify two things, above all: which tasks you are expected to comply with in this seminar, and how your grade is going to be composed based on this.

2nd appointment, 26th april, 1st round of play:

> After being handed out the instructions for the game, you will gain first experience with playing the game in randomly allotted groups. Furthermore, you will be randomly divided into groups of 4 in which you will have the opportunity to successively develop a set of rules for the game situation during the weeks to come. For simplicity, we will call these sets of rules "virtual rooms" (vRoom). Please appoint a contact person for your group by 3rd May, whose contact details are given to us for possible enquiry calls.
>
> Important note: You will be randomly allotted to a vRoom in each round of play, and you will not know with whom you are interacting in a vRoom in any of the rounds.

3rd appointment, 3rd May, 1st closing date:

> Handing in the rule set by midnight of the respective day; rules are to be handed in including a documentation in which you shortly (max. 2 pages) expose your deliberations on the set of rules handed in.

4th appointment, 10th May, 2nd round of play (single-weightet):

First-time play in the newly-designed vRooms; you will, again, be randomly assigned to your respective vRoom, and therefore, you cannot count on playing with the same people as the week before, nor with other members of your rule group.

5th appointment, 17th May, Ascension Day; handing-in of the 2nd rule set

6th appointment, 24th May, 3rd round of play (single-weighted)

7th appointment, 31st May, handing-in of the 3rd rule set

8th appointment, 7th June, 4th round of play (double-weighted)

9th appointment, 14th June, handing-in of the 4th rule set

10th appointment, 21st June, 5th round of play (final round) (triple-weighted)

11th appointment, 12th July, final discussion in the large group

12th appointment, 31st July, handing-in of the field reports

Your tasks:

- participating on all appointed days is an absolute necessity for this seminar, as otherwise there cannot be any play in at least one vRoom! In case of severe illnesses, we ask for the earliest possible notification!
- active participation in 5 rounds of play
- active participation in designing a rule set for the game situation. Additionally, a short motivation for your rule set. (Group task)
- preparation of a seminar paper in which you analyse your experience made in the seminar and in which you concludingly evaluate your rule set. (Individual task)

Composition of your grade:

- seminar paper and documentation of the rule sets: 40%
- individual play: 30%
- performance of your group's rule set in the efficiency tournament: 20%
- final discussion: 10%

The efficiency tournament of rule sets (vRooms):

Efficiency will be defined as follows: the sum of individual payoffs within the vRoom minus rule-set costs as a fraction of the sum of maximum possible payoffs without rule set. I.e., if players "do very well" under your rule set, you will earn points in the efficiency tournament of vRooms. In order not to give you any incentive to play "worse" in vRooms of others, your individual payoffs from the round of play are also taken into account for your grade.

Note, furthermore, that the efficiency that emerges in your vRoom during rounds of play 2 and 3 will contribute to your overall score in the efficiency tournament single-weightedly, while efficiency gained in rounds 4 and 5 will be weighted twice and three times, respectively. This is not the case for individual payoffs.

What is a "rule set", what does its implementation cost, and which rule sets are feasible?

To start with, a "rule set" is everything that could be introduced to induce the players to perform actions that increase the total sum of contributions in the corresponding vRoom. Exceptions are technically unfeasible changes (or those that are feasible only under prohibitively large costs, as e.g. to transfer the laboratory to the attic, providing it with a glass cupola), as well as ethically questionable practices (shoot other players, cut their hands off). Additionally, rules are excluded, that make inferences about the true identity of players feasible.

As in "real life", there is one thing to think about: "there ain't no such thing as a free lunch" – every rule to be introduced gives rise to specific costs that will be taxed by us according to the difficulties the introduction/implementation of the rule would entail in "real life".

In addition to the dates appointed above, each group should arrange with Mr Wolff for a consultation hour on the Tuesday or Wednesday preceding each handing-in of the rule sets to discuss the costs of any potential rule set and to find out whether specific rule changes are not possible due to technical, ethical, or other reasons, and to thus be able to change the rule set accordingly in this case. It is up to you to decide on how many group members go to these consultation meetings.

A2. Instructions for the basic game

Instructions for the experiment

General Information:

At the beginning you will be randomly assigned to **3 groups of 4 participants**. You will not be informed about the identity of the other group members.

Course of Action:

In every round, you will be given an endowment of 20 points you can invest in a common project. You have to decide how many of the 20 tokens you are going to contribute to the project. You will keep the remaining tokens.

Calculation of your payoff in stage 1:

Your period payoff consists of two components:

tokens you have kept = endowment – your contribution to the project

earnings from the project = 1.6 x sum of the contributions of all group members / number of group members

Thus, **your period payoff** amounts to:

20 – your contribution to the project
+ 1.6 x sum of the contributions of all group members / number of group members

The earnings from the project are calculated according to this formula for each group member. The total payoff from the experiment is composed of the sum of period payoffs from all 25 rounds. Payoff scores will remain anonymous, i.e. no participant will be informed of the payoff score of any other participant.

Please notice:

Communication is not allowed during the whole experiment. If you have a question please raise your hand out of the cabin. We will then come to you and answer your question privately.

Good luck!

Table B1: summary of all individual rule sets with elicited contribution levels, achieved efficiency and costs, according to seminar, group, and round of design.

Seminar	**Grp**	**Rule set 1**	**Contributions**	**Efficiency**	**Costs (variable costs)**	**Rule set 2**	**Contributions**	**Efficiency**	**Costs (variable costs)**
1	1	- endogenous state authority - punishment - endogenous redistribution - ID	825	2277	218 (158)	- (lagged) punishment after warning	668	1747	654 (454)
	2	- trial rounds - punishment - redistribution - feedback: indiv. contributions	550	1968	362 (62)	- redistribution of rule costs - no feedback (but own payoff)	641	2085	300 (0)
	3	- punishment - ID	789	1625	848 (648)	- roundly varying background stories - individual messages	533	2270	50 (0)
2	4	- punishment - individual messages	1437	2302	560 (310)	- punishment - redistribution - individual messages - big bonus	1795	2822	255 (5)
	5	- punishment - redistribution	361	1915	302 (102)	- minimum-effort - appeal	1725	2583	452 (252)
	6	- punishment - appeal	597	1829	530 (280)	- no feedback - 'big malus' for minimum-contributor if efficiency < 2500 (before rule-costs) - appeal	1271	2263	500 (0)

Seminar	Grp	Rule set 3	Contributions	Efficiency	Costs (variable costs)	Rule set 4	Contributions	Efficiency	Costs (variable costs)
1	1	- redistribution - ID	734	2150	290 (90)	- redistribution + redistribution of rule-costs - ID	1449	2661	208 (8)
	2	- (endogenous) redistribution of rule-costs administered by high-contributors - ID	627	2076	300 (0)	- redistribution of rule costs (every 4th round) - vote on institution	978	2486.0	101 (0)
	3	- individual messages (changing every round)	703	2372	50 (0)	- cheap-talk contract (full contribution; no longer available after 3rd breach) - roundly appeals - individual feedback	1411	2796.6	50 (0)
2	4	- punishment - redistribution - individual messages - big bonus	1091	2291	364 (114)	- redistribution - big bonus - example in instructions (msg)	2000	2850.0	350 (0)
	5	- minimum-effort - punishment - redistribution - appeal	2000	3000	200 (0)	- cheap-talk contract offered every 5th round: endogenous 'minimum-contribution' - individual messages	418	2200.8	50 (0)
	6	- no feedback - 'big malus' for minimum-contributor if efficiency < 3000 (before rule-costs)	1547	1828	1100 (600)	- appeals every 5th round	390	2184.0	50 (0)

Legend (fixed rule costs):

appeal (50)

A statement is displayed to subjects before they play the game. The statement is possibly repeated every *k* rounds. Usually, the statement would appeal to moral values and/or a group sentiment.

big bonus (0)*

In every playing round, points are deducted from the players conditional on their contribution behavior and transferred into a 'jackpot account' applying the formula for redistribution. At the end of the 25 rounds, the 'jackpot' is awarded to the highest-contributing player, with an equal-split rule in case of a tie.

'big malus' (100)*

The lowest-contributing player is punished by 200 points if the group does not achieve a certain efficiency benchmark. Costs are calculated as if the player is deducted 8 points each round (Total efficiency cost in case of application: 200 + 400 = 600 points).

cheap-talk contract (50)

Players are given the opportunity to vote on a 'contract' requiring them to contribute a specified amount of tokens. The contract is 'enacted' only in case of unanimous consensus. The 'enactment' does not have any material consequences. In case of the endogenous 'minimum-contribution' levels, players were meant to vote on increasing the initially specified level of 15 tokens by one token in case of full compliance.

endogenous state authority (0)

Individual contributions are disclosed if the group invests 4 points into a secondary public good 'administration'; individual contributions to this 'administration' cannot surpass 2 points. If the 'administration' is 'constituted', its contributors may engage in peer-to-peer punishment (cost schedule as in *punishment*).

example in instructions

An example specifying the payoffs of a single free-rider amongst full-contributors is contrasted to the payoff of a full contributor amongst equals.

ID (200)

Players are given fixed ID numbers and informed of individual contributions of all players by ID number.

individual messages (50)*

Pre-defined messages, automatically displayed to players conditional on their behavior in the game.

minimum-effort (0)*

'Redistributing' back 'surplus contributions': any tokens contributed by a player *i*, which surpass the smallest contribution made by a group member are returned to *i*, diminished according to the redistribution formula, before the sum of contributions is multiplied by the public-good factor.

no feedback (but own payoff; 0)

Players do not receive any feedback (except on their own payoff), not even about the sum of contributions.

punishment (0)*

Deduction of x points, $x < 9$, from a player's income from the public good according to the formula:
$Cost(x) = x^2/4$. These costs are born by all players if punishment is administered automatically, and by the punishing player in case of peer-to-peer punishment. Higher deductions are possible, but no perfect punishment automata are available: in this case, players either have to administer punishment themselves, or non-contributions are punished only with a certain probability. The corresponding automata have additional fixed costs:
High probability punishment (probability of deduction = 1/2): 600 points
Low probability punishment (probability of deduction = 1/10): 200 points

(lagged) punishment after warning (0)*

A warning is issued to a player who does not comply with a specified condition. In case of repeated non-compliance, the player is

	punished automatically as described under *punishment*.
(endogenous) redistribution (0)*	Transfer of points from one player to another/others. The points available for allocation, y, as a function of points deducted, x, are determined according to the following formula: $y = sqrt(x - 1) + 1$. In the case of endogenous redistribution, individual players (determined in correspondence to their contributions) may decide on the reallocation of points.
redistribution of rule costs (0)*	In contrast to the default case, rule costs are allocated unequally among players. Under rule sets of this category, players are told they have to bear different shares of the rule costs depending on their contribution behavior.
roundly varying background stories (50)	The public good is framed in a different way in each round. An example for a framing would read: "In your neighborhood, cats are poisoned more and more often. Residents of the area are worried that the rat poison laid-out may also be eaten by dogs or even by little children. They are forming a vigilante group in charge of combing through the streets for irregularities and to collect the poison. As you do not have the time to participate in these activities, your family is thinking about donating € 20 for support. Do you want that?", followed by a yes-or-no decision. In case of a "no", the story goes on: "It could be your little daughter who swallows such a toxic bait. And do you want to lock up your cat the whole day long? The vigilante group is a good thing, how much do you want to give them? It's for the safety of your family?"
trial rounds (100)	Players experience the rule set for a small number of rounds without payoff consequences before they play the game for 25 rounds.

*Rule features marked are feasible only in conjunction with monitoring; a monitor requires costs of 200 points, independent of whether she announces the observed contributions as in *ID* or not.

THURGAU INSTITUTE
OF ECONOMICS
at the University of Konstanz

Hauptstr. 90
CH-8280 Kreuzlingen 2

Telefon: +41 (0)71 677 05 10
Telefax: +41 (0)71 677 05 11

info@twi-kreuzlingen.ch
www.twi-kreuzlingen.ch