Groups and scores: the decline of cooperation

Stefano Duca[1,*] **and Heinrich H. Nax**[1]

[1]ETH Zürich, D-GESS, 8092 Zürich , Switzerland
[*]sduca@ethz.ch

ABSTRACT

Cooperation between unrelated individuals in social-dilemma-type situations has been a focus of many studies in social and biological sciences. It has repeatedly been shown that, without suitable mechanisms, high levels of cooperation/contributions in repeated public goods games cannot be stable in the long run. Reputation, as a driver of indirect reciprocity, is often proposed a mechanism that leads to cooperation. A very prominent reputation dynamic functions through scoring: contributing behavior increases one's score, non-contributing reduces it. Indeed, many experiments have established that scoring can sustain cooperation in two-player prisoner's dilemmas and donation games. However, these prior studies focused on pairwise interactions, with no experiments studying reputation mechanisms in more general group interactions. In this paper, we focus on groups and scores, proposing several scoring rules that could apply to multi-player prisoners' dilemmas played in groups, which we test in a laboratory experiment. Results are unambiguously negative: we observe a steady decline of cooperation for every tested scoring mechanism. All scoring systems suffer from it in much the same way. We conclude that the positive results obtained by scoring in pairwise interaction do not apply to multi-player prisoner's dilemmas, and that alternative mechanisms need to be considered.

Introduction

Social dilemmas are situations where the optimal decision from the perspective of a self-interested individual conflicts with what is optimal for the group collectively. Public goods[1] and common-pool resources[2], as modeled using game theory via prisoner's dilemmas (PD), voluntary contributions games[3,4] or donation games[5] all constitute examples of social dilemma situations. The common feature of these interactions is that it is known that in the absence of a suitable mechanism[6,7] and/or sufficient foresight by the players[8,9], the only stable outcome that is achieved coincides with the socially undesirable one: absence of cooperation/contributions/donations, etc.[1]. The players fail to cooperate and, as a result, they are all worse-off than in the collective optimum. This phenomenon is often referred to as the "tragedy of the commons"[10,11] or the "free riders problem"[12].

One of the most important mechanisms that successfully implements cooperation is "indirect reciprocity"[13,14]. Indirect reciprocity is a behavioral pattern/norm whereby people will return benefits for benefits (and hostility with hostility)[15]. Cooperation will breed cooperation and lead to higher payoffs. Thus, cooperation overcomes the momentary benefit of defection, which breeds defection and eventually leads to low payoffs. A principal driver of indirect reciprocity is reputation[16], therefore known as a "universal currency"[17]: cooperating, or refusing to do so and choosing to defect, not only affects one's stage-game payoff but also impacts one's reputation. When interacting again in the future, players will take each others' reputations into account, thus indirectly reciprocating as players cooperate only with those who have good reputation (i.e. have contributed in the past). This creates incentives to cooperate beyond the momentary temptation of defection, provided the future benefits of cooperation are substantial. As a result, cooperation through indirect reciprocity typically emerges in the presence of reputation mechanisms.

Indeed, reputation –via numerous implementations– has been shown to stabilize high levels of cooperative behavior in controlled human experiments[18,19]. An important limitation of prior studies has been that they all focus on pairwise interactions. In reality, however, most human social interactions unfold in groups[20]. This is particularly relevant in present society as interactions increasingly take place online, involving largely impersonal, crowd interactions and partially anonymous features.

Moving from pairwise interactions to group interactions substantially complicates matters, in theory and in practice. In a group interaction, players might not be able to observe the actions undertaken by others and are unable to infer individual actions from aggregrate outcomes, thus making it harder to track and update other players' reputations. Other than in a two-person interaction, one can often not infer the others' individual actions from one's own action and the outcome. For instance, when playing a public good game, information regarding individual behaviors may not be available, and the only available behavioral information may concern the group as whole.

[1]For the sake of clarity, we will use cooperate as a common terminology for all terms like cooperate, contribute, and donate.

This raises the following key question. More generally, how do reputation mechanisms fare in group interactions? More specifically, as the general question is too broad to be addressed here, we focus on what is one of the best-known and perhaps simplest mechanisms to implement reputation: "scoring". Our analysis of scoring in groups extends the concept of "image scoring"[21,22], as has been studied widely in pairwise interactions. In a basic image scoring[23] scenario, each player has a score (starting at 0) that represents his reputation. Whenever a player has the opportunity to cooperate with someone else (alternatively, to be kind or help someone else), his score gets updated: if he cooperates with the receiver his score is increased by one, if not it is decreased by one. Thus a player's reputation is continuously reassessed based on past decisions (in the simplest case based on the previous decision). A seminal theory result[23] is that the strategy to cooperate with anybody with a non-negative image score is evolutionary stable. Crucially, by refusing to cooperate with/help someone with a low image score a player is decreasing his own score, thus reducing his own probability of receiving cooperation/help in the future. Hence, not cooperating with a player with low image score can be interpreted as a form of punishment. Indeed, in practice, numerous behavioral experiments show that image scoring helps stabilize cooperative behavior in two-player PDs and donation games[22,24–26].

As we extend scoring mechanisms to group interactions more generally, and to multi-player PDs in particular, we increase the degree of freedom regarding the scoring rules that may be said to apply. Real-world group interactions vary with respect to the information that is available, and typically individuals do not observe all the actions undertaken by all the other individuals, especially in large groups. The relevant scoring mechanism that applies to a specific group interaction will therefore depend on how much information is available to players, and on how much information each reputation rule will require. Processing available information correctly may become difficult in group interactions. Indeed, it was conjectured[17] that the reason for why image scoring is favored over other reputation dynamics is that (relatively) little information is required to implement image scoring in two-player interactions with full feedback[27]. Investigating cases involving limited feedback[28] or feedback with errors[29] has revealed that sufficient accurate information was key for mechanism success, and that cooperation may break down behaviorally in these cases. When interacting in groups, due to increased anonymity of players, information becomes coarser. A single subject may thus find it harder to reap the benefits of cooperative "reputation-building" strategies, and cooperation may break down for the same reason.

Recent theory has extended "scoring" methods to group interactions[30]. The baseline establishes a positive cooperation result for the case of image scoring in group interactions. Extending this result, group interactions are considered where only information regarding group performance –but not regarding individual players– is available. In such situations, "group scores" replace image scores: each player's group score summarizes the aggregate cooperativeness of the groups to which he belonged in the past, without any additional information regarding what players individually did. In this case, theory predicts that cooperation cannot be sustained at all.

To date, there has been no study of behavioral data concerning scoring mechanisms and indirect reciprocity in group interactions. It is an open challenge to understand how reputation matters in a group setting depending on informational context. In this paper, as a first step toward addressing this question more generally, we investigate whether simple implementation variants of scoring mechanisms will sustain cooperation in multi-players PDs. We conducted a laboratory experiment to test candidate scoring rules that would apply in group interactions with different information availability. The baseline is to test image scoring. In addition, we propose alternative scoring rules that could apply to group interactions including a scoring mechanism where players score each other through votes. The proposed rules differ in how much information regarding past behavior of their group-mates is required, ranging from no feedback to full feedback, thus implementing a version of image scoring in a group setting.

The experimental results concerning cooperation are negative: for every tested scoring mechanism we observe a steady decline in cooperative behavior. The decay of cooperation is the same under every mechanism and comparable even with the case when no scoring mechanism at all is implemented. We conclude that positive results regarding cooperation deriving from scoring, as were repeatedly observed in two-player interactions, do not generalize to group interactions. We establish this for the case of multi-player prisoner's dilemmas, contrary to theoretical predictions regarding image scoring in groups.

The rest of this paper is structured as follows. Next, we present the experimental results obtained for the different scoring mechanisms. Finally, we discuss our results. Method sections contain details about the experiment and about the statistical analyses.

Results

Before presenting results, we briefly discuss the structure of the experiment and introduce the different scoring mechanisms that were tested. For further detail concerning the experimental design, we refer the reader to the Methods section.

Experimental procedure

A large experiment involving 192 subjects playing multi-player PDs and making a total of 11,520 decisions was performed. The experiment was conducted in 12 sessions involving 16 subjects each. Each session of the experiment consisted of three different treatments that were played over 20 rounds each. In each treatment, subjects were faced with a different scoring rule/mechanism. Every round, subjects were randomly re-matched in groups of size 4 and provided with feedback about the score of their group-mates from the previous round, as calculated using the current scoring rule. After deciding whether to contribute or not, subjects received individual payoffs (of which they were informed) and were assigned updated scores. It is important to note that, by virtue of our design, the score of a subject only reflected his last action, and that scores did not carry over multiple rounds of the game. Our focus is on situations where mechanisms are introduced or where a new mechanism replaces an old one. Hence, subjects in our experiments always initially played a treatment were no feedback about others' actions or scores were given. After that, two different scoring mechanisms were played in succession.

Scoring mechanisms and hypotheses

We tested the following scoring mechanisms ranging between image scoring and no scoring at all:

- **No scoring:** Subjects receive no information at all regarding the past actions of the other players, and therefore it is the treatment with the lowest informational content. *Hypothesis:* In this implementation of a repeated multi-player PD we expect a decay of contribution resulting in low contribution levels, as shown by numerous previous experiments[22,26,27,31] mainly conducted in voluntary contribution games settings.

- **Image scoring:** This is the treatment with the highest informational content of all, equivalent to the case with a binary image score in two-players interactions. Players are told whether their past and future group-mates contributed to the common pool in the previous round. *Hypothesis:* Based on previous experiments on donation games[31] and on theoretical results[30], one could expect a stable high level of contributions.

- **Group scoring:** Scoring proceeds as in image scoring, except that all group members receive the same score based on the number of contributors in their group. Subjects are given no direct information about individual decisions. *Hypothesis:* Recent theoretical work[30] suggests that a low level of cooperative behavior is to be expected.

- **Self scoring:** Players directly assign the score to their fellow players based on feedback regarding own payoffs and aggregate contributions in their group. This treatment potentially contains more information than the group scoring but of course players might not be truthful when assigning the scores. *Hypothesis:* In this case the only Nash equilibrium is for nobody to contribute, independently of the assigned ratings.

- **Truthful self scoring:** This is a control treatment for self scoring, where scores are exogenously assigned reflecting the true actions of the players; meaning that the scores are automatically assigned as if all the players were truthful in the Self scoring treatment. It is important to note that the informational content here is, in principle, equivalent to Image scoring, but it might be harder for players to interpret this formulation.

Experimental results

In fig. 1 we show the percentage of contributors as a function of time for all the different treatments.[2] The figure on the left shows the contribution levels observed during the second phase of the experiment, and the one to the right the ones observed during the third phase. As first treatment (i.e. in the first phase) subjects always played the treatment with no scoring. For all the different treatments, we observe a steady decline in average contributions. The decay occurs in much the same way and it seems to be independent from the order in which the different treatments were played.

Even though it is possible to statistically distinguish some of the observed downward trends from some others (e.g. image scoring is significantly different from group scoring, see table 1), the main difference in treatments can be reduced to a slight offset in the initial percentage of contributors. Fig 2 illustrates that the estimated (linear) decay of cooperation over time occurs at the same speed. Indeed, all the slopes are within the error range of each other. Differences are noticeable only regarding the intercept, that is, regarding initial contributions.

For more details on the statistical analysis, we refer the reader to the Methods section.

The above results clearly indicate that even scoring mechanisms that were shown to stabilize high level of cooperation in two-players games (i.e. image scoring), fail to achieve the same results when multi-player interactions are concerned. The most plausible explanation is that it is harder to isolate the "bad apples" from the crowd in group interactions, resulting in a deterioration in the quality of information, as perceived by subjects. This lack of targeting precision could be all that it takes to destabilize cooperative equilibria: To keep stable high levels of cooperation, in fact, players should –on average– cooperate

[2] More detailed plots are available in the supplementary materials.

with a frequency at least as high as the observed number of players with a high score in their group, thus maintaining a stable percentage of "good players" in the population. We instead observed[3] that, while ceteris paribus players do cooperate more with an increased observed score in their group, they do so with a downward bias, especially for high sums of scores in the group. Further contributing to this phenomena is the fact that when a high-score player decides not to cooperate because of the presence of low-score subjects in his group, it might reduce the score of all his group-mates, not just of the low-score individuals. This results in a steady shrinking of players with good reputation in the population and consequentially in the break down of the positive effect of scoring on cooperative behaviour observed in pairwise interactions. The arising effect could be considered akin to the downward spiral of contributions in time observed by Fischbacher et al.[38] when studying conditional cooperation in a public good experiment.

Discussion

Scoring methods in general, and image scoring in particular, are simple implementations of reputation mechanisms. They stabilize cooperative behavior in various standard, two-players social dilemma situations, such as in prisoner dilemmas or donation games. Image scoring requires reliable feedback regarding individual-level behavior as input. The purpose of this study is to extend such mechanisms to group interactions, in particular to multi-players prisoner dilemmas where individual-level data could be hard to obtain. We propose several scoring rules that apply in this setting, depending on informational context, and tested them in a laboratory experiment. Furthermore, we also test how an endogenized scoring mechanism could be implemented. The results are unambiguously negative: independent of information, feedback and scoring mechanism, cooperation decays. This includes mechanisms that were shown to stabilize cooperation in the corresponding two-player case. A plausible explanation is that individuals cannot be isolated; i.e. defectors cannot be individually punished, and cooperators cannot be individually rewarded. This results in a reaction to the average score in the group biased toward defection, leading to a steady decrease of high reputation players in the population that in turn begets lower level of cooperative behaviour.

On a broader level, our results show that there is still much that we do not know about reputation dynamics. Even though indirect reciprocity is considered one of the main mechanisms trough which cooperation can be sustained, there have been very few studies on interactions in group setting. Understanding such settings has become particularly relevant in recent years because, due to the increasing digitalization of our world, more and more interactions take place online where people frequently communicate via crowd platforms and where often explicit reputation tallying is provided as a method to build trust. Due to the increasing decentralization of interactions, partial or total anonymity of the actors involved can be the norm and reputation is often build on a peer to peer basis with members of communities rating each other. For example, a project may involve several groups of individuals and information on individual level contributions could be imperfectly filtered via several community-layers before reaching the players. With this work we set up to investigate some of these issues. The initial results show that many positive results on cooperation, repeatedly observed in pairwise interactions, do not hold anymore when groups are concerned.

Many facets of this issue should be subjects of future work, for example: are dynamics in play in multi-player games fundamentally different from the ones in two-player games? And if so, could one exploit this to device a scoring mechanism able to sustain higher levels of cooperation? Future work should address such issues and many others. Further experiments on the dynamics of multi-player games might be needed to guide future theoretical research and suitably address such a fundamental aspect of human behaviour.

Methods

The experiment

The experiment was conducted at the ETH Decision Science Laboratory in Zurich and we used the z-Tree[39] software to design and run the experiment. We ran 12 sessions with 16 participants in each session, for a total of 192 participants. Players were recruited using the hroot[40] sofware and were mainly university students. Each session in the laboratory lasted roughly one hour during which the players played 3 treatments for 20 rounds each.

Subjects always played first the treatment were no information regarding past behavior was provided. After that, subjects played two other treatments. In the table below we detail the 6 different treatments' combinations that were played in two separate sessions each.

At the beginning of the session and before each treatment, subjects were given written instructions[4] explaining what the experiment was about and the game that they were about to play, scoring mechanism included. Before the first treatment, subjects were given some minutes to familiarize with the game with a small training. Before the Truthful self scoring treatment,

[3] See supplementary material for details.

[4] Available in the supplementary material.

	Treatment combinations **Round 1-20 → Round 21-40 → Round 41-60**
Control	*No scoring → No scoring → No scoring*
Treat. Com. 1	*No scoring → Image scoring → Group scoring*
Treat. Com. 2	*No scoring → Group scoring → Image scoring*
Treat. Com. 3	*No scoring → Image scoring → Self scoring*
Treat. Com. 4	*No scoring → Image scoring → Self scoring*
Treat. Com. 5	*No scoring → Self scoring → Image scoring*

because of the complexity of the scoring mechanism, subjects also had some minutes to understand how the scoring worked using a score simulator.

As customary, subjects were incentivized by converting their earnings in real currency. Subjects on average earned 33 CHF (roughly 33 USD), including 5 CHF of show-up fee. Earnings ranged from 25 to 40 CHF.

In the following we define the game that the subjects played in the experiment and the scoring mechanisms that were used.

N-players prisoner's dilemma

The subject played the following game whose aspects were all common knowledge:

1. For 20 rounds, subjects are randomly assigned to separate groups of fixed size 4.

2. Subjects decide whether to contribute their endowment to the common pool or not.

3. Subjects receive individual payoff ϕ according to $\phi_i = (1 - c_i) + \frac{1}{2}\sum_{j=1}^{4} c_j$, where $c_i = 1$ if subjects i contributed, and 0 if not.

4. *Scoring:* a score is assigned to each player subject on his contribution (depending on the treatment). The score is visible to the other subjects in the following round, and subjects learn whom they will be grouped with next.

Regardless of the treatment, all subjects were shown the profit made during the round and during the entire session; thus each subject was told how many people contributed in his group in the previous round. It is important to notice that the score of a subject only reflected its last action and it was not carried over multiple rounds of the game.

The scoring mechanisms

Here we explain in detail how, depending on the treatment, the score was assigned to each subject. In a given round of the game, subjects were shown the scores (if any) assigned in the previous round. For all treatments, the score ranged between 0 and 1.

- **No scoring:** No score at all was assigned to players during this treatment.

- **Image scoring:** Subjects were assigned a score of 1 if they contributed in the previous round and 0 if not.

- **Group scoring:** Subjects were assigned a score proportional to how many people in their group contributed to the common pool. The score was given by the number of contributors in their group, divided by the group size (4) and thus subjects in the same group all received the same score. Therefore, the higher subject i's score, the higher is the probability that i invested in the group account. Please note that in case of a score of 0 or 1, the group score faultlessly indicates whether a subject contributed or not.

- **Self scoring:** Each subject was asked to rate his/her group awarding a number of stars ranging from 0 to 3. The score of each subject was computed as the sum of all the stars awarded to the group by his group-mates (excluding his own rating) divided by 9. Hence, the score of each subject could rank between 0 (all his group-mates awarded 0 stars to the group) to 1 (all his group-mates awarded 3 stars to the group).

- **Truthful self scoring:** The score was assigned as in the self scoring treatment but exogenously. This means that each subject was considered as having awarded a number of star to his group equivalent to the number of contributors (excluding himself) observed in his group.

Statistical Analysis

To find out if treatments significantly differ from one another, we used the Mann-Whitney-Wilcoxon rank sum test[41,42]. Let us call ${}^{q}x_i^t \in \{0,1\}$ the decision that player i took at time t in treatment q. We first compute the average decision in a treatment at each time step: ${}^{q}\overline{x}^t \equiv \sum_{i=1}^{16} {}^{q}x_i^t$, with ${}^{q}\overline{x}^t \in [0,1]$. We define the "time series" ${}^{q}\overline{x} \equiv \left\{{}^{q}\overline{x}^1,...,{}^{q}\overline{x}^{20}\right\}$. After that, we perform a rank sum test for each pair of treatments. The p-value obtained from the rank sum test is a measure of how likely it is that ${}^{i}\overline{x}$ and ${}^{j}\overline{x}$ are drawn from the same distribution. Table 1 shows the p-values for each pair of treatment in the second phase of the experiment. A value depicted in red indicates that the two treatments statistically significantly differ from each other.

To obtain fig. 2 we performed a linear regression of the contributions to the public good as a function of time for each treatment individually and for all of them together. Further statistical analysis is provided in the supplementary material.

References

1. Isaac, R. M., McCue, K. F. & Plott, C. R. Public goods provision in an experimental environment. *Journal of Public Economics* **26**, 51–74 (1985).
2. Ostrom, E. *Governing the commons: the evolution of institutions for collective action* (Cambridge University Press, 1990).
3. Andreoni, J. Why free ride? strategies and learning in public goods experiments. *Journal of Public Economics* **37**, 291–304 (1988).
4. Isaac, R. M. & Walker, J. M. Group size effects in public goods provision: the voluntary contributions mechanism. *The Quarterly Journal of Economics* 179–199 (1988).
5. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).
6. Riolo, R. L., Cohen, M. D. & Axelrod, R. Evolution of cooperation without reciprocity. *Nature* **414**, 441–443 (2001).
7. Hetzer, M. & Sornette, D. A theory of evolution, fairness, and altruistic punishment. *PloS ONE* **8**, e77041 (2011).
8. Osborne, M. J. & Rubinstein, A. *A course in game theory* (MIT press, 1994).
9. Friedman, J. W. A non-cooperative equilibrium for supergames. *The Review of Economic Studies* **38**, 1–12 (1971).
10. Ostrom, E. Coping with tragedies of the commons. *Annual Review of Political Science* **2**, 493–535 (1999).
11. Hardin, G. The tragedy of the commons. *Science* **1243** (1968).
12. Baumol, W. J. *Welfare economics and the theory of the state* (Longmans, Green and co, London, 1952).
13. Alexander, R. D. *The biology of moral systems* (Transaction Publishers, 1987).
14. Mailath, G. J. & Samuelson, L. *Repeated games and reputations: long-run relationships* (Oxford University Press, 2006).
15. Gouldner, A. W. The norm of reciprocity: a preliminary statement. *American Sociological Review* 161–178 (1960).
16. Ostrom, E. A behavioral approach to the rational choice theory of collective action: Presidential address, american political science association, 1997. *American Political Science Review* **92**, 1–22 (1998).
17. Milinski, M. Reputation, a universal currency for human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150100 (2016).
18. Panchanathan, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**, 499–502 (2004).
19. Fehr, E. Human behaviour: don't lose your reputation. *Nature* **432**, 449–450 (2004).
20. Perc, M., Gómez-Gardeñes, J., Szolnoki, A., Floría, L. M. & Moreno, Y. Evolutionary dynamics of group interactions on structured populations: a review. *Journal of The Royal Society Interface* **10**, 20120997 (2013).
21. Nowak, M. A. & Sigmund, K. The dynamics of indirect reciprocity. *Journal of Theoretical Biology* **194**, 561–574 (1998).
22. Wedekind, C. & Milinski, M. Cooperation through image scoring in humans. *Science* **288**, 850–852 (2000).
23. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
24. Wedekind, C. & Braithwaite, V. a. The long-term benefits of human generosity in indirect reciprocity. *Current Biology* **12**, 1012–1015 (2002).
25. Semmann, D., Krambeck, H. J. & Milinski, M. Reputation is valuable within and outside one's own social group. *Behavioral Ecology and Sociobiology* **57**, 611–616 (2005).
26. Seinen, I. & Schram, A. Social status and group norms: indirect reciprocity in a repeated helping experiment. *European Economic Review* **50**, 581–602 (2006).

27. Milinski, M., Semmann, D., Bakker, T. C. & Krambeck, H.-J. Cooperation through indirect reciprocity: image scoring or standing strategy? *Proceedings of the Royal Society of London B: Biological Sciences* **268**, 2495–2501 (2001).

28. Bolton, G. E., Katok, E. & Ockenfels, A. Cooperation among strangers with limited information about reputation. *Journal of Public Economics* **89**, 1457–1468 (2005).

29. Berger, U. & Grüne, A. On the stability of cooperation under indirect reciprocity with first-order information. *Games and Economic Behavior* (2016).

30. Nax, H. H., Perc, M., Szolnoki, A. & Helbing, D. Stability of cooperation under image scoring in group interactions. *Scientific Reports* **5** (2015).

31. Milinski, M., Semmann, D. & Krambeck, H.-J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426 (2002).

32. Saijo, T. & Nakamura, H. The "spite" dilemma in voluntary contribution mechanism experiments. *Journal of Conflict Resolution* **39**, 535–560 (1995).

33. García, J. & Traulsen, A. Leaving the loners alone: evolution of cooperation in the presence of antisocial punishment. *Journal of Theoretical Biology* **307**, 168–173 (2012).

34. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367 (2008).

35. Rand, D. G., Armao IV, J. J., Nakamaru, M. & Ohtsuki, H. Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology* **265**, 624–632 (2010).

36. Rand, D. G. & Nowak, M. A. The evolution of antisocial punishment in optional public goods games. *Nature Communications* **2**, 434 (2011).

37. Chaudhuri, A. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics* **14**, 47–83 (2011).

38. Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters* **71**, 397–404 (2001).

39. Fischbacher, U. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* **10**, 171–178 (2007).

40. Bock, O., Baetge, I. & Nicklisch, A. hroot: Hamburg registration and organization online tool. *European Economic Review* **71**, 117–120 (2014).

41. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of Mathematical Statistics* 50–60 (1947).

42. Richardson, A. M. Nonparametric statistics: a step-by-step approach. *International Statistical Review* **83**, 163–164 (2015).

Acknowledgements

We thank Claus Wedekind, Manfred Milinski, Urs Fischbacher and his team at TWI Kreuzlingen, Stefan Wehrli, Oliver Brägger, Jakob Dambon, and our colleagues from COSS at ETH Zurich for helpful discussions and comments. All remaining errors are ours. Financial support from the European Commission through the ERC Advanced Investigator Grant 'Momentum' (Grant 324247) is gratefully acknowledged.

Author contributions statement

S.D. and H.H.N. conceived the experiment, conducted the experiment, analysed the results, and reviewed the manuscript. S.D. coded the experiment.

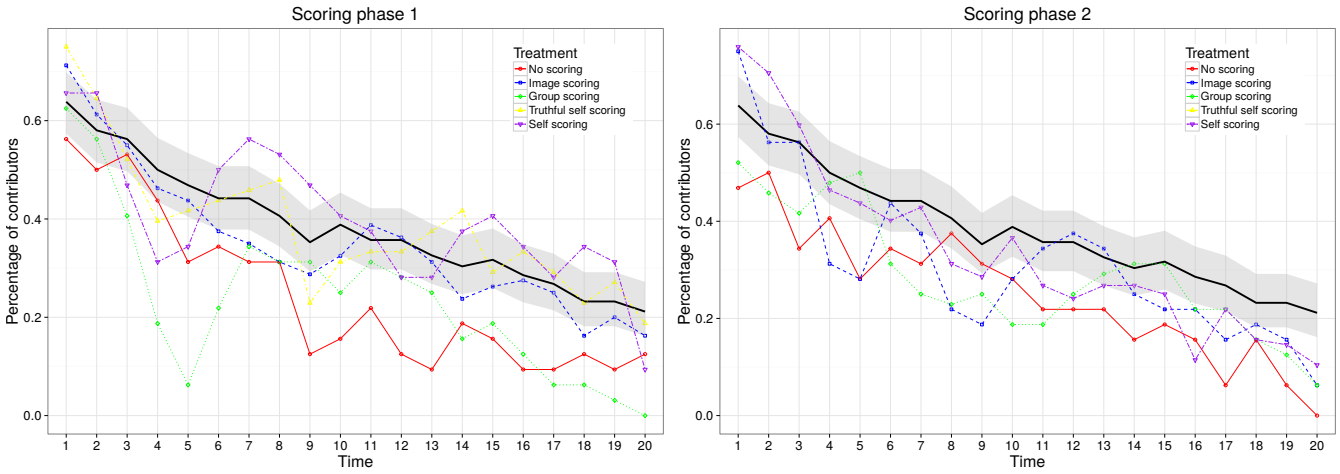
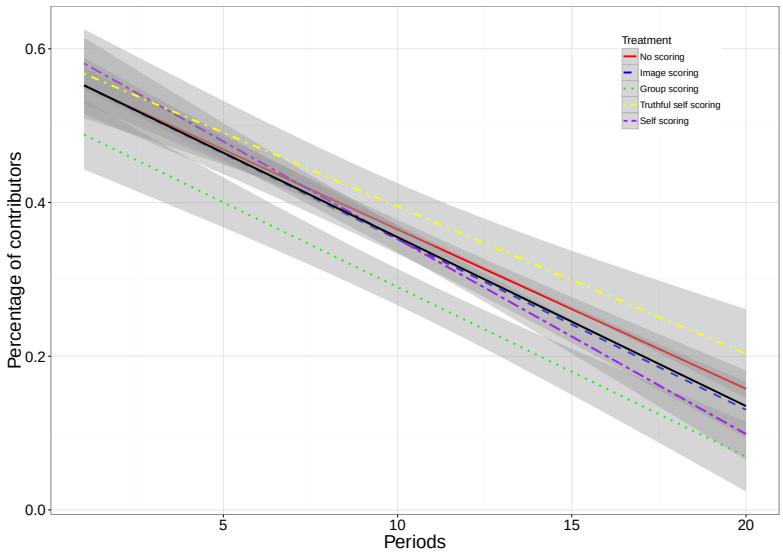


Figure 1. Percentage of contributors as a function of time for all the treatments: regardless of the treatment we observe a great decline in cooperative behaviour over time. The figure on the left shows the contribution levels observed during the second phase of the experiment and the one to the right the ones observed during the third phase. The black line in the background shows the average contribution level observed in the first phase. As first treatment (i.e. in the first phase) subjects always played the treatment with no scoring and hence it can be treated as a baseline. The grey area represent the binomial proportion confidence interval. The figures clearly show a great decline in average contributions. The decline seems to happen in much the same way for all the different treatments and it seems to be independent from the order in which the different treatments were played.

	No scoring	Image scoring	Group scoring	Truthful self scoring	Self scoring
No scoring	n/a	0.156	0.903	0.006	0.005
Image scoring	0.156	n/a	0.019	0.350	0.155
Group scoring	0.903	0.019	n/a	0.003	0.001
Truthful self scoring	0.006	0.350	0.003	n/a	0.542
Self scoring	0.005	0.155	0.001	0.542	n/a

Table 1. Mann–Whitney–Wilcoxon Rank Sum test for each pair of treatment. The table shows the p-values obtained from the Mann–Whitney–Wilcoxon Rank Sum test for each pair of treatment in the second phase of the experiment. A p-value depicted in red means that the time series obtained from the two treatments are likely drawn from two different distributions. A black p-value indicates that the time series obtained from the two treatments might be drawn from the same distribution. A red p-value indicates that it is likely that the treatments come from two different distributions.



(a)

Treatment	Estimated slope
No scoring	-0.020 ± 0.001
Image scoring	-0.022 ± 0.002
Group scoring	-0.022 ± 0.002
Truthful self scoring	-0.019 ± 0.003
Self scoring	-0.025 ± 0.002
All data	-0.022 ± 0.001

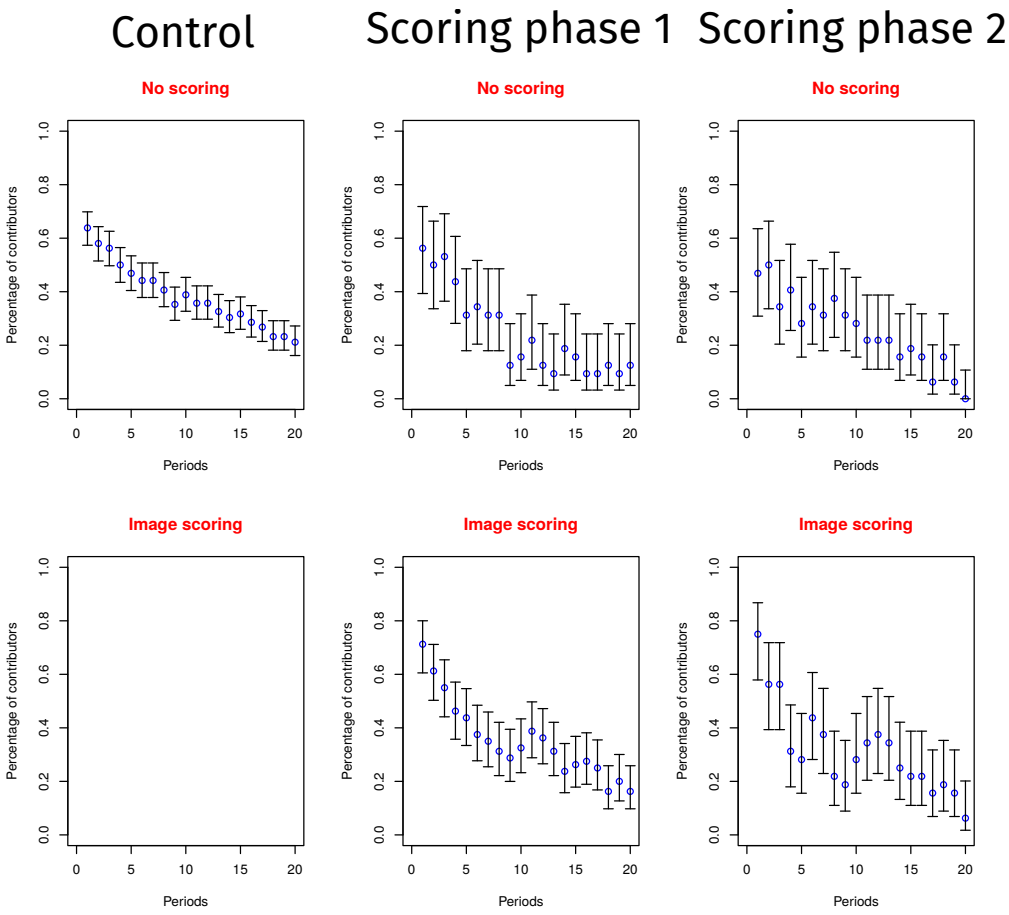
(b)

Figure 2. Estimated linear function representing the decay of the percentage of contributors as a function of time: all treatments decay in the same way. In figure (a), each coloured line illustrates the fitted linear function for a different treatment. The grey areas depict the 95% confidence interval. The black line depicts the estimated decay for the entire data set. Table (b) lists the values obtained for the various slopes from the linear regression. As we can observe, there is a difference in the intercept of the different lines but all treatments decline with the same slope. Indeed, all slope values are well within the error range of each other, indicating no difference in rates of decay.

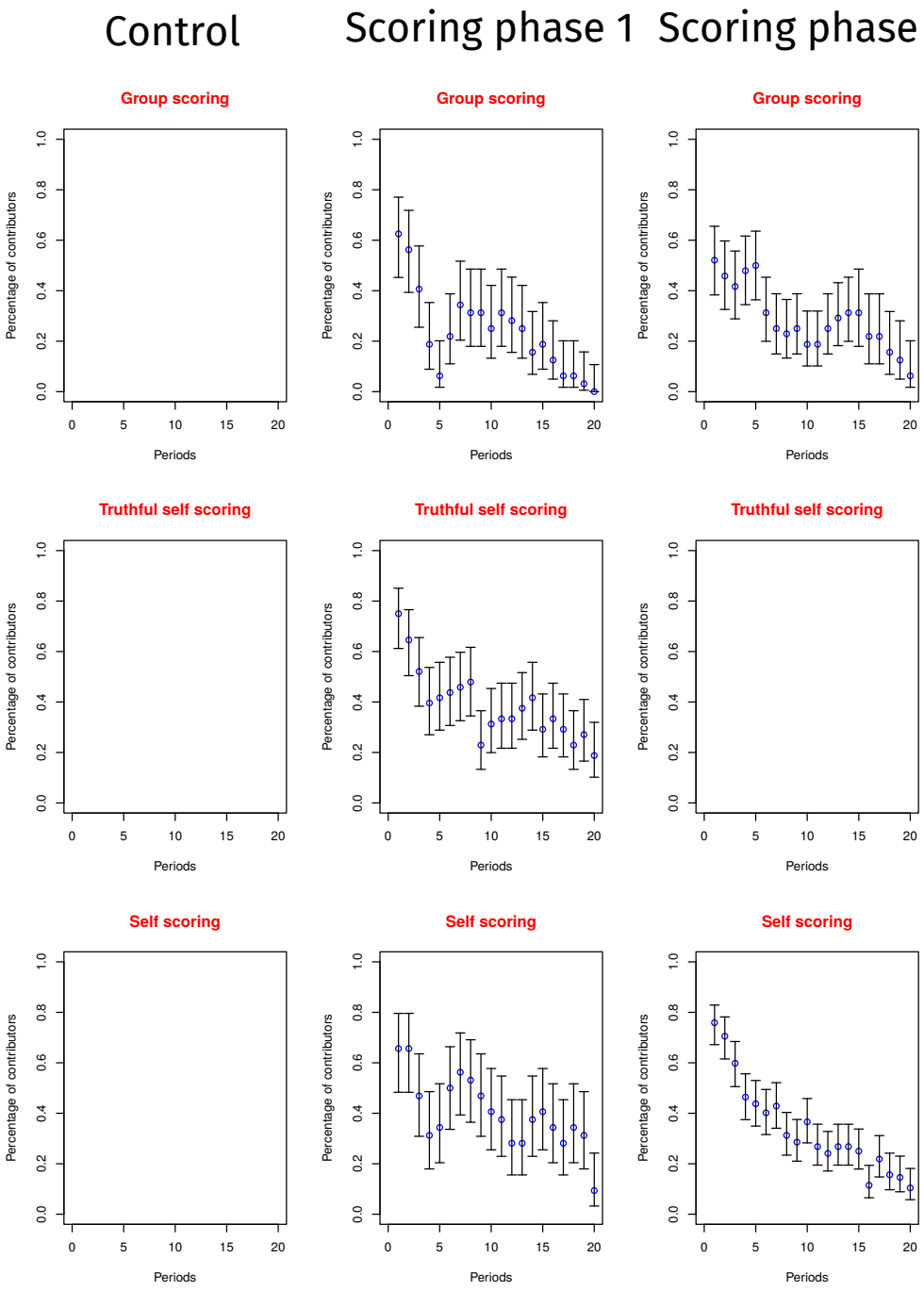
Supplementary Material

Detailed experimental results

In the following we show the percentage of contributors as a function of time for each of the treatments and phases. The bars indicate the 95% Binomial proportion confidence interval[1]. The difference in the error bars' size depends on the wide difference in the number of available data points.



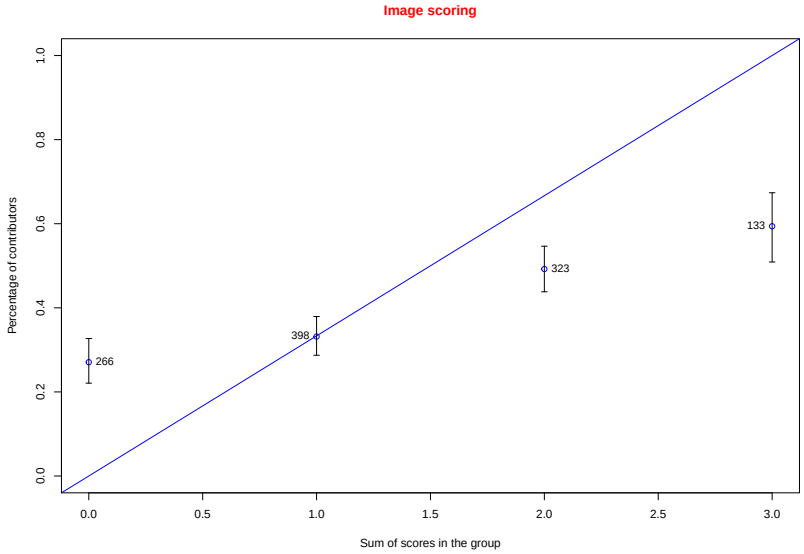
[1]For an exact definition see e.g. Brown, Lawrence D., T. Tony Cai, and Anirban DasGupta. *Interval estimation for a binomial proportion.* **Statistical science** (2001): 101-117.

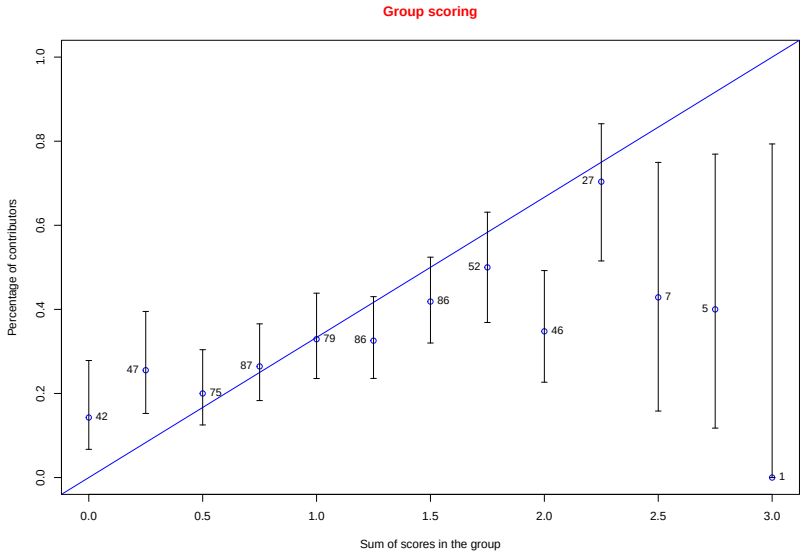

Control
Scoring phase 1
Scoring phase 2
Group scoring
Group scoring
Group scoring
Truthful self scoring
Truthful self scoring
Truthful self scoring
Self scoring
Self scoring
Self scoring
Percentage of contributors
Periods
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20

Percentage of contributors as a function of the observed score in their group

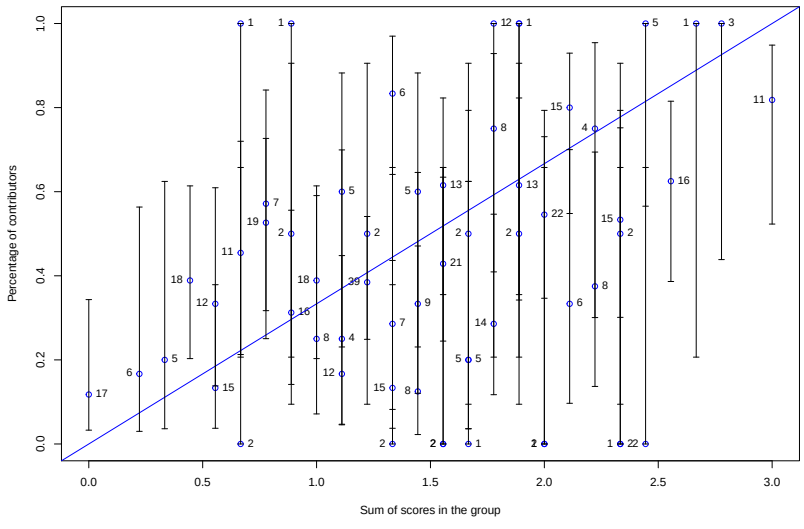
In the following, we plot the percentage of contributors as a function of the observed score in their group. The bars indicate the 95% Binomial proportion confidence interval[1]. The difference in the number of points in the figures comes from the fact that different scoring mechanisms produce a different number of possible combinations of the scores of the players in a group.

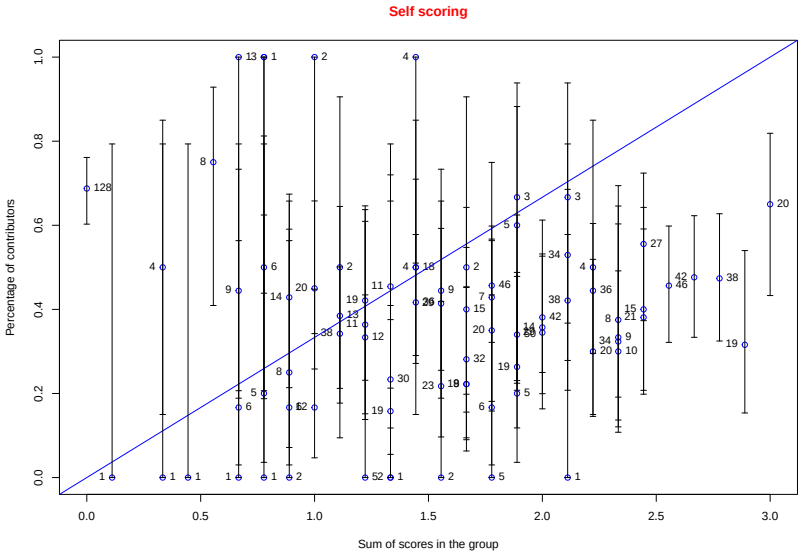
In the Image and Group score pictures, we can observe that players seem to contribute more with increasing observed score in their group. However, players seem to do so with a downward bias, especially for high score values.





Due to the high numbers of possible values for the Self and Truthful Self scores, the related figures are more chaotic but a similar trend seems to be present there as well.




Self scoring
Percentage of contributors
Sum of scores in the group
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Further statistical analysis

Because of the potential correlations between decisions taken by player i at time t and the decisions taken at $t-1$ by all the other players, we decided to perform another statistical test that does not rely on computing the average decision at each time step ${}^q\bar{x}^t$.

For this reason, we performed a permutation test using the between sum of squares (SSB) as statistics:

$$SSB = \sum_{i=1}^{m} n_i \left(\bar{y} - \bar{y}_i\right)^2$$

The idea behind the test is to randomly permute treatments within sessions many times, thus creating a random sample of all the possible matches between our observable and the treatment under which they have been observed (see figure for a two dimensional example with 3 treatments tested 3 times each).

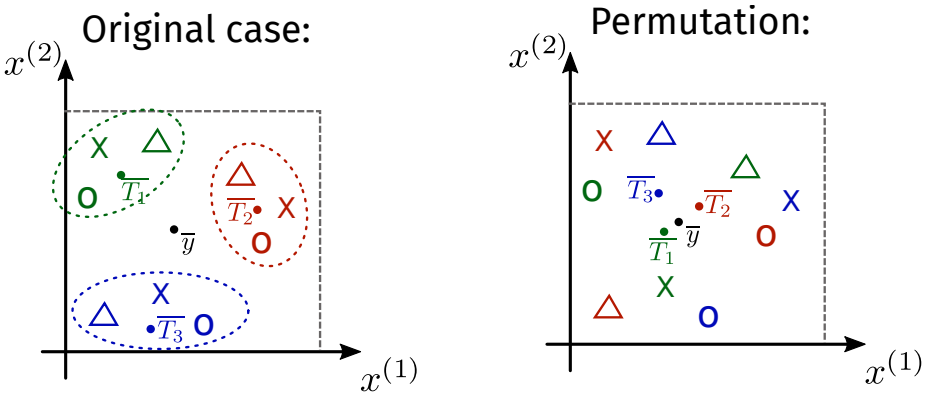


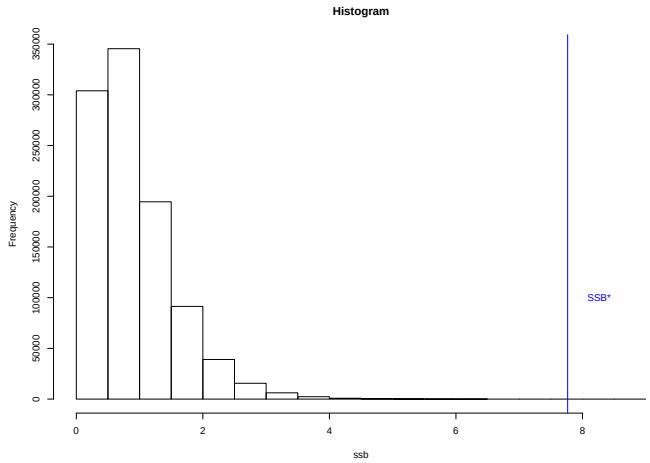
Figure 1: Example of permutation. Every symbol represents data collected in an experiment session. X, △ and O indicate that the data were collected during the first, second and third session respectively. Green indicates that data were collected during treatment 1, red during treatment 2 and blue during treatment 3. $\overline{T_i}$ is the mean for treatment i while $\overline{y}$ represents the global mean.

For every random permutation we computed the resulting SSB, thus obtaining an histogram representing the empirical distribution of the SSB values.

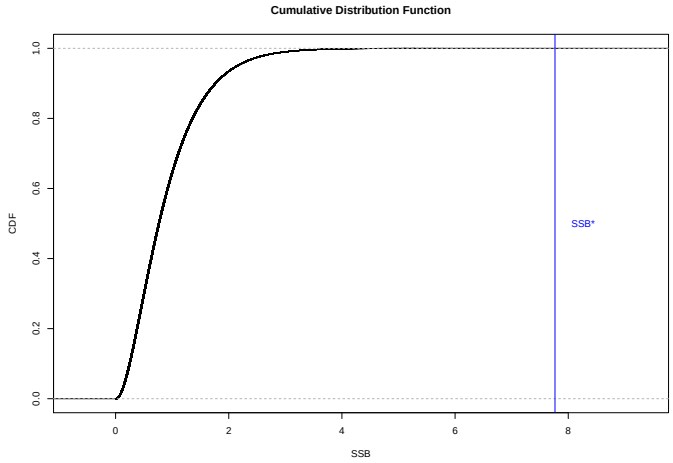
We can now compare two hypothesis: The null hypothesis (H_0), where treatments have no impact on the decisions taken by the agents and the alternative hypothesis (H_1) where we should observe statistical significance of treatments.

To compare the two hypothesis, we calculate the SBB for our original data (let us call it SBB^*) and we compute how likely it is (using the empirical observed cumulative distribution function) to observe SBB^* under H_0.

Here is what we get:



(a) The histogram represents the frequency of observed sbb values obtained from 1000000 permutations. The blue line indicates where the SBB for the original data lies.



(b) Here we plot the empirical cumulative distribution function obtained from the permutation. From it, it is possible to compute the empirical p-value for SBB^*.

Figure 2: The figures show where the

From the figures above we can clearly observe that H_0 is rejected and thus that the treatments are not all statistically undistinguishable from each other (the observed p-value is< 0.001). This is in line with the results presented in table 1 where we observe that some treatments are significantly different from each other.