Document de travail du LEM / Discussion paper LEM 2019-03 "Version révisée"

Group formation and cooperation in social dilemmas: A survey and meta-analytic evidence

Andrea GUIDO

Catholic University of Lille, ETHICS and LEM-UMR CNRS / andrea.guido@univ-catholille.fr

Andrea ROBBETT

Middlebury College, Department of Economics / arobbett@middlebury.edu

Rustam ROMANIUC

LEM UMR 9221 / rustam.romaniuc@gmail.com

http://lem.cnrs.fr/

Group formation and cooperation in social dilemmas: A survey and meta-analytic evidence*

Andrea Guido†, Andrea Robbett‡& Rustam Romaniuc§

December 19, 2018

Abstract

We survey the growing literature on group formation in the context of three types of social dilemma games: public goods games, common pool resources, and the prisoner's dilemma. The 62 surveyed papers study the effect of different sorting mechanisms – endogenous, endogenous with the option to play the game, and exogenous – on cooperation rates. Our survey shows that cooperators are highly sensitive to the presence of free-riders, independently of the sorting mechanism. We complement the survey with a meta-analysis showing no difference in terms of cooperation between studies implementing an endogenous and exogenous sorting. What is more, we find that it is no more likely for a cooperator to be matched with like-minded partners in endogenously formed groups than in exogenously formed groups. These observations are related. As we show in the survey, the success of a sorting method in matching like-minded individuals and the levels of cooperation are closely interlinked.

*For valuable insights at various stages of this work, we wish to thank Rachel Croson, Fabrice Le Lec, Marie-Claire Villeval and the participants at the IMEBESS meeting in Barcelona, the ASFEE meeting in Rennes, and at the Cooperation: Interdisciplinary Methods Workshop in Lille. We are also grateful to the authors who responded to our various requests on the ESA forum.

†Corresponding author. Catholic University of Lille, ETHICS and LEM-UMR CNRS. Contact: *andrea.guido@univ-catholille.fr*

‡Middlebury College, Department of Economics. Contact: *arobbett@middlebury.edu*

§Catholic University of Lille, ETHICS and LEM-UMR CNRS. Contact: *rustam.romaniuc@gmail.com*

1 Introduction

There is overwhelming evidence from laboratory and field experiments that genetically unrelated individuals cooperate even when this is individually costly and does not provide benefits in the form of delayed gratification (Rand and Nowak [2013]). However, an equally frequent observation is that cooperation declines over time because some people switch to free-riding when they observe selfish behavior in their group. In other words, research in experimental economics demonstrates that the co-existence of "conditional cooperators" and "free-riders" may lead to socially undesirable outcomes over time (e.g. Ehrhart and Keser [1999], Keser and van Winden [2000], Burlando and Guala [2005], Gächter and Thöni [2005], de Oliveira et al. [2015]).

In the last two decades, many experiments manipulated the way groups are formed in order to find a solution to the decay of cooperation over time. Chaudhuri [2011] notes that group formation is one of the most effective non-punitive mechanisms to sustain cooperation. The literature started with the seminal work by Ehrhart and Keser [1999] where the authors used the wind-tunnel of a laboratory experiment to test Tiebout [1956]'s principle of endogenous sorting. The papers that followed explored two research avenues: (i) variations to the endogenous sorting mechanism, and (ii) the effect of exogenous sorting on cooperation.

We survey the growing literature on group formation in the context of three types of social dilemma games that implement different sorting mechanisms: public goods games, common pool resources, and the prisoner's dilemma. The 62 surveyed papers implement three types of sorting:

1. endogenous sorting: all subjects play the social dilemma game where subjects have some ability to form or leave groups on their own accord;

2. endogenous sorting with the option to play the game: subjects can exit a specific relationship or avoid the game entirely;

3. exogenous sorting: sorting is undertaken by the experimenter with or without the subjects' knowledge.

From the surveyed literature, we observe that, independently of the sorting mechanism, cooperators are highly sensitive to the presence of free-riders in their group and therefore, the specific composition of the groups created play a crucial role in sustaining high cooperation levels over time. We identify the design specifications that result in groups composed of cooperative types. In experiments with an endogenous sorting, we observe that the

possibility to engage in costly signaling of one's cooperative disposition results in a separating equilibrium. Similarly, in games where participants can opt not to play (either at all or temporarily), there is a selection effect: cooperators have higher expectations about the likelihood that others will cooperate and are more likely to opt-in to a social dilemma. The downside of these mechanisms of generating groups is that free-riders do not learn to cooperate. A different strand, within the endogenous sorting literature, allows for costly punishment of free-riders, which helps sustain cooperation over time. What is more, with the presence of punishment, free-riders switch to cooperative behavior over time and groups reach homogeneity in behavioral types despite differences in intrinsic predispositions to cooperate. Additionally, the composition of the generated groups has been put forward as the main feature that allows cooperation to evolve in experiments with exogenous sorting. In games that use a one-shot interaction to sort out subjects, the separation of cooperative types from free-riders appears to be more successful than in games that use repeated interactions or multiple methods to generate groups. The presence of punishment is less effective at increasing cooperation in exogenously-formed groups. In fact, punishment and exogenous sorting may be substitutes rather than complements. As noted by Gächter and Thöni [2005], punishment does not have a significant effect on contributions in either exogenously-formed groups of high cooperators or exogenously-formed groups of low cooperators and therefore plays less of a role when groups have already been sorted.

There is substantial difference in the experimental protocols between endogenous and exogenous sorting studies, which makes it difficult to state whether one leads to higher levels of cooperation than the other based solely on a review of the literature. From a theoretical point of view, we should expect no difference in contributions between the two sorting mechanisms. Indeed, an informed experimenter could induce the same sorting exogenously as under endogenous sorting. However, if the sorting method is imperfect in some way when groups are created, endogenously or exogenously, then the generated groups may differ in their type-composition between the two sorting mechanisms, leading to differences in cooperation rates. Additionally, there is robust experimental evidence showing that endogenous institutional choices encourage cooperation compared to a situation in which the same institution is imposed by the experimenter (Sutter et al. [2010], Dal Bo et al. [2010], DeAngelo et al. [Forthcoming]). If free-choice – or some form of control aversion – matters in the context of group formation, we should observe higher contributions in the endogenous sorting compared to the exogenous for a similar type-composition of the generated groups. We conduct a meta-analysis to investigate empirically the effect of the two sorting

mechanisms on cooperation controlling for some environmental and design variables. The results suggest that there is no significant difference in contributions between studies with endogenous sorting and studies with exogenous sorting. The data, additionally, confirm that it is no more likely for a cooperator to end up in a group composed mostly of co-operators under endogenous sorting than under exogenous matching. As we show in the survey sections, the success of a sorting method in matching like-minded individuals and the levels of cooperation are closely interlinked. By examining the probability of being matched with like-minded others, we highlight one of the main reasons for the absence of difference in cooperation levels between endogenous and exogenous sorting.

Our contribution to the literature on the determinants of cooperation is twofold. First, we provide researchers in experimental economics with an overview of the methods that have been used to create groups endogenously and exogenously and, most important, highlight their advantages and disadvantages in terms of achieved levels of cooperation. Second, the literature on group formation has always considered separately how the two sorting mechanisms affect cooperation. This is the first study to compare the effects of endogenous sorting and exogenous matching on cooperation.

The remainder of the paper is organized as follows: section 2 presents the endogenous sorting literature, section 3 presents the exogenous sorting papers, in section 4 we present the selection criteria and the main results for the meta-study, and section 5 concludes.

2 Endogenous sorting

Chaudhuri [2011] defines endogenous sorting as the mechanism that allows subjects to enter or exit groups on their own accord. The endogenous sorting mechanism has been studied in more than 40 laboratory experiments. The objective of this section is to identify the different ways through which endogenous sorting mechanisms are implemented and their effects on cooperation.

We will begin by presenting the simplest mechanism that allows subjects to freely enter and exit groups – hereafter, *free migration*. Papers in this category are rooted in Charles Tiebout's canonical local public finance model Tiebout [1956], which proposed that residents will "vote with their feet" in response to differences in local public good provision and move to the communities where the public good and other local features best match their preferences. Ehrhart and Keser [1999] conducted the first experiment testing Tiebout's hypothesis. In their experiment, subjects are allowed to move or create

new groups.[1] The game consists of a 30-round linear public goods game with the marginal per capita return (MPCR) from the public account decreasing in the group size.[2] In each session, 18 subjects play two different independent games of 9 subjects each.

The authors find that the composition of groups is not *stable* and dynamics are *non-monotonic*: highly cooperative groups in the current period tend to grow in following rounds, while those characterized by low contributions shrink. However, when the size of the group remains unchanged or increases, contribution levels unravel afterwards. This unstable dynamic is due to free-riders trying to enter groups of cooperative individuals who, in turn, leave to create new groups.

Given the diversity in individuals' preferences to cooperate, the parameters chosen by Ehrhart and Keser [1999] imply one efficient outcome that has not been reached in the experiment: the segregation of individuals with high dispositions to cooperate. However, the formation of homogeneous, segregated groups may not always be optimal. Robbett [2016] studies the dynamics of community formation when members have different preferences – i.e., different marginal per capita returns from the public good. The population is split into two different types: High types who benefit more from the provision of the good and may experience congestion and Low types who have a low marginal per capita return.[3] Six groups or locations are formed at the beginning of the experiment and subjects can move across them. Two different games are implemented: a congestible public goods (CPG) game, in which the presence of free-riders reduces High types' payoffs (as in Ehrhart and Keser [1999]) and a pure public goods (PPG) game, where the most efficient outcome is for all participants to form a single group and High types have no financial incentives to move to avoid the Low types. The results show that heterogeneity in preferences is a relevant factor to understand the migration patterns and subjects' contributions. High types are generally the first to enter empty communities and they contribute high amounts over time. Low Types enter highly cooperative groups later, contribute less, and the overall contribution rates decline over time. While there is a substantial difference in the migration patterns between the CPG and the PPG games, the congestion drives only about half of the movement. High types continue to flee Low types even when their presence does not lower their financial payoff, suggesting that much of the chasing pattern observed by

[1] Moving across groups is costly. Subjects are endowed with 10 Experimental Currency Units (ECUs) and had to pay 5 ECUs for creating or switching groups.

[2] In the standard linear public goods game, each participant is endowed with tokens or currency that they can allocate between a private account (which benefits the individual) and a public account (which benefits all group members). The marginal return from the private account is typically 1, while the per-person return from the public account, the MPCR, is less than one.

[3] Agents are aware of the heterogeneity but not of the distribution.

Ehrhart and Keser [1999] and in the CPG game is driven by an intrinsic preference to avoid free-riders.

These papers therefore suggest that cooperation may only be prevented from unraveling by screening and separating defectors from conditionally cooperative individuals. The introduction of a costly signal – for example a higher group entry fee – could allow pro-social agents to stay away from free-riders who would not be willing to incur the high-cost of faking the signal.

2.1 Endogenous sorting with signaling opportunities

Brekke et al. [2011] aim at testing whether social commitments, such as charity donations, could serve as a costly screening device. In their 3-person public goods game, subjects can choose to adhere to one of the two group types: in the *blue* group, each member receives an extra fixed payoff while in the *red* group the extra sum is donated to the Red Cross. The experiment is divided into three parts: part I consists of a one-shot public goods game with groups formed randomly;[4] in part II subjects play 10 rounds of the public goods game choosing beforehand their preferred group type; part III is the same as part II with the difference that players can change groups in each round and the number of rounds is increased to 20. The authors expect more cooperative subjects to join the red group since socially beneficial commitments can serve as costly screening devices to help pro-social subjects form homogeneous groups.

Results show that higher contributors in the one-shot game choose the red group while those who contribute less prefer the blue one. This correlation between willingness to cooperate and choice of group demonstrates the validity of the signal. As a consequence of the successful sorting, cooperation levels in red groups are higher and stable over time compared to those of blue groups, which present the usual decreasing trend.[5] Furthermore, when Brekke et al. [2011] run a placebo treatment where subjects in the red groups received no extra money. They observe that few subjects, if any, chose the red group and cooperation unraveled as there was no separation between types. The validity of the signal seems strictly related to its *social meaning*.

Other experiments have demonstrated the effectiveness of endogenous mechanisms without reliance on social meaning. For instance, Aimone et al. [2013] design a labora-

[4]The subjects were informed that the experiment consists of 3 parts and that choices made in part I will not affect their earnings in the two other parts of the experiment.

[5]In a follow-up study, Hauge et al. [2018] explore whether subjects classified as altruists, free-riders, conditional cooperators and others differ in their propensity to self-select into red groups. They find that more generous subjects self-select into red groups, while the less generous tend to choose the blue groups.

tory experiment, absent any group identity or doctrinal construct.[6] In their experimental setting, subjects are endogenously sorted by revealing their willingness to relinquish a fraction of their private return from the public good. The experiment considers two one-shot games: a standard and a "sacrifice" public goods game. The latter differs from the former in that subjects can express their preferences regarding the group's private return by choosing one value within [0.55, 0.95] in 0.05 increments, while it is fixed at 1 in the standard game. After all subjects express their preferences, the four subjects with the highest private return are grouped together, the four subjects with the next highest private return are placed together in another group, and so on. The private return chosen in each group is the average number chosen by the 4 group members. Two treatment orderings are used: the normal public good game in the first round followed by the "sacrifice" in the second (Inexperienced treatment) and vice versa (Experienced treatment). In both orderings, the Sacrifice game is played at the end of the experiment. Subjects are aware of the assortative mechanism and unaware whether there would be a subsequent round in each stage.

Results show that more cooperative individuals choose a lower private return (higher unproductive cost), thus isolating themselves from defectors, regardless of the game-play order. To further investigate the causes of the sorting mechanism's success, three control sessions with an exogenous (random) grouping mechanism were run. From this comparison, it is clear that the successful provision of the public good is driven by the costly sacrifice that signals one's cooperative disposition, which is impossible in random exogenous mechanisms.

Signaling mechanisms have been tested even under uncertain conditions regarding their validity. Grimm and Mengel [2009] implement a prisoner's dilemma game introducing the concept of *viscosity*: an increased probability of interacting with others of one's type or group. Subjects can choose to play in groups that differ in the defector gain of the associated payoff scheme, thus signaling one's disposition to cooperate. However, a low viscosity reduces the benefit of the signal as it is likely that subjects might be paired with others from the opposite group. Subjects play the game for 100 rounds, repeatedly choosing between two groups: group A, whose defector gain in the related payoff scheme is lower than that of group B. They conduct three treatments that differ in their viscosity levels and three control conditions (varying the feedback and payoffs given to subjects and ruling out the possibility of sorting). Prior to playing the game, subjects were informed about the percentage of subjects in groups A and B, and their individual probability of meeting members of each group.

[6] The authors base their work on Iannaccone [1992] related to "sacrifice and stigma" mechanisms – i.e., unproductive costs, employed by religious groups to dodge free-riders.

The data suggest that the level of viscosity affects the sorting of subjects into different groups. In treatments characterized by high levels of viscosity a large fraction of subjects choose group A (35%-60%) and cooperate most of the time, while subjects in the other group almost never cooperate.[7]

In the long run, the separating equilibrium, achieved in some of the afore-mentioned papers, implies the absence of interactions between free-riders and cooperators. The segregation of cooperative subjects, albeit beneficial for themselves, does not compel free-riders to change their behavior. To do that, the design of the experiment needs to allow subjects to *enforce* cooperative social norms, sanctioning socially inappropriate behavior and rewarding socially appropriate conduct. We distinguish three main enforcement mechanisms: monetary punishment, ostracism, and entry and exit restrictions.

2.2 Endogenous sorting with punishment

Gürerk et al. [2006]'s paper is one of the first to investigate the role of punishment under free migration.[8] The authors aim to address the following question: would a sanctioning institution deliberately be adopted when individuals can choose between a sanctioning and a sanctioning-free environment? In their experiment, subjects play a 30-round public goods game. Each round has three stages: an institution choice stage, a voluntary contribution stage and a sanctioning stage. In the first stage, subjects choose between joining a group with the presence of punishment and reward (SI) or one without (SFI). In the second stage, all the participants interact with subjects in their own group in a public goods game. Eventually, in the third stage, participants grouped in SI are asked to reward or sanction other members of the same group by assigning from -20 to 20 tokens to other participants. Each token negatively assigned induces a cost of 3 ECU to the punished subject and 1 ECU to the punisher. Conversely, in case of tokens positively assigned, the cost and the reward are symmetrically set at 1 ECU. Subjects receive feedback regarding other participants' individual contributions from the same group as well as from the other groups.

One third of the population chooses SI in the first round. The initial choice of the institution correlates with different types of behavior. More than half of those who choose

[7] The literature on endogenous group formation where subjects get payoff-relevant information about the others (or about other groups) includes Bayer [2016], Bohnet and Kübler [2005], Coricelli et al. [2004], Cabrera et al. [2013]

[8] Page et al. [2005] is an earlier and related paper investigating the role of punishment. In their 2x2 design, Page et al. [2005] compare an endogenous sorting mechanism with (and without) punishment to a baseline with randomly created groups. They find that punishment significantly improves cooperation. The major difference with Gürerk et al. [2006] is that in their experiment people do not migrate across different institutional environments.

SI in the very first round are classified as "high-contributors". Three-quarters of them use tokens to punish and establish norms of cooperation. This fraction of subjects can be classified as "strong reciprocators". Only 5% of subjects sorted in the SI group are classified as free-riders (with contributions lower than 5 ECUs), which is not surprising because opting for a sanction and not contributing would be an example of shooting oneself in the foot. This fraction hikes up to 44% in the SFI group, revealing free-riders' preference for a sanction-free environment.

Free-riders in the SFI environment initially earn more and therefore many subjects join this group starting from the second round. As cooperation unravels in the immediate periods, subjects find it optimal to migrate towards SI where payoffs are higher. Subjects switching from SFI to SI significantly increase their contributions and even former free-riders become full cooperators. Cooperation emerges as a generally accepted norm in SI and subjects use punishment only occasionally over time.[9]

In a follow-up paper, Gürerk [2013] investigates whether *social learning* can increase the willingness to accept punishment institutions. The experimental design is similar to Gürerk et al. [2006] with the difference that rewards are not considered. Three treatments are implemented: a baseline game where subjects receive no information regarding previous sessions, a treatment (SHT) where participants receive a report about the decisions made by participants of a previous experiment (see Gürerk et al. [2010]) and a treatment (SHT-Half) where subjects are provided with only a subset of the social history. The social history provided consists of the average number of community members, their contributions, the received punishment tokens, and the payoff for each round in a treatment with punishment implemented by Gürerk et al. [2010]. In the SHT-Half treatment, only the history of the institutional choice is provided to subjects.

It turns out that social history increases the initial acceptance of the punishment institution and achieves full participation in the community with punishment. Contributions steadily increase when social history is provided and stabilize near the social optimum. This in turn lowers punishment expenses and reduces the initial efficiency loss seen in Gürerk et al. [2006]. Furthermore, the significant difference between SHT and SHT-Half leads to the conclusion that subjects do not merely imitate the institutional choice made by other subjects but also pay attention to other relevant factors.

The experiments by Gurerk and co-authors implement decentralized, peer-to-peer pun-

[9]Rewards are not perceived as encouragement to increase contributions, as they target those who already abide by the social norm of cooperation. Gürerk et al. [2014] disentangle the effect of rewards from punishment in a follow-up work.

ishment. However, the external validity of these results is weakened if one thinks that modern societies are permeated with centralized punishment. Nicklisch et al. [2016] experimentally investigate the formation of groups when agents can choose between a sanctioning-free environment, an environment with decentralized punishment, and an environment with centralized punishment. In the centralized punishment environment, the authority meting-out the sanction is randomly chosen among the participants to the experiment. Subjects can switch groups every 4 periods and receive feedback concerning others contributions to the group account. In only one treatment the information about other subjects' individual contribution is perfectly accurate. In the remaining two treatments, the information is accurate at 90% and 50% of probability. Hence, players can receive a distorted information about others' contributions. The authors expect that the accuracy of information given to subjects will significantly affect their preferences as to the punishment environment. In particular, highly noisy information should discourage the adoption of any form of punishment.

Their results suggest that subjects are equally attracted to all three institutions when information is accurate at 90%, while decentralized punishment is the majority's choice in sessions with perfect information. As expected, the majority of subjects choose the sanctioning-free institution when information is very noisy (accuracy at 50%) because under these conditions monetary punishment may be highly detrimental. The other dark side of punishment is that it creates social losses. Many experimental studies have focused on testing other forms of enforcement. Cinyabuguma et al. [2005] use a public goods game to study the influence of ostracism on cooperation. In each session, groups of 16 subjects are endowed with 10 ECUs and play a series of 15-round public goods games. The experiment has two treatments: the *baseline* treatment consists in a standard public goods game, while in the *Expulsion* treatment subjects can vote to expel members of their own group after being informed of each others' past contributions. The expulsion is implemented only if the majority votes in favor. The ostracized subjects are banished to another group (called the "Blue group") playing the same public goods game with a reduced endowment of 5 ECUs for the remaining rounds. Each subject voting in favor of expulsion faces a cost of 25 ECUs in case of a majority vote to expel the targeted subject. The authors implement two conditions: BE where subjects play the Expulsion treatment after the Baseline, and EE where subjects play the Expulsion twice.

The authors find that expulsion is used sporadically, on average between one and four times per session in each treatment. The related dynamic fluctuates over time: it peaks

in the first round where ostracism is used to initially discipline defectors and in later rounds due to the end-game effect when cooperation unravels. Votes are targeted towards defectors. Targeted players immediately raise their contribution levels after having faced the threat of expulsion.[10] When ostracism is possible, contributions are higher starting from the first round because subjects anticipate this punishment, and are stable over time until the end of the game. In the EE condition, contributions are higher the second time the expulsion treatment is played than the first time.

Maier-Rigaud et al. [2010] find similar results in an experimental setting involving groups of 6 subjects playing a 10-period public goods game with a partner matching design. They differ from Cinyabuguma et al. [2005] in that they do not allow ostracized players to play in a group of likewise expelled players. Instead, they are simply excluded from all group activities, earning their endowment in each round. They also implement smaller group sizes and consequently a higher MPCR, and the vote to expel other group members is costless. Two treatments characterize the design of this experiment: a baseline consisting in a simple public goods game, and a treatment with the presence of ostracism.[11]

Ostracism has a positive effect on contributions, particularly if included in the first treatment of the session.[12] Despite differences in the experimental settings, results are similar to Cinyabuguma et al. [2005]. Average contributions in the baseline and the ostracism treatment significantly differ in all periods, apart from the first rounds when the social norm of cooperation has not been well established yet, and last rounds of each treatment, with the end-game effect.[13] However, forms of antisocial punishment also emerge over time, with high contributors being the target of defectors' votes.

In some experimental settings, participants can not only ostracize group members but also freely migrate across groups and/or be forgiven. In Charness and Yang [2014]'s experiment, 9 subjects are randomly assigned to 3 groups and play a 3-player public goods game for three periods. At the end of it, subjects learn about IDs and individual contributions of their own group members. The authors then allow subjects to exit their groups, to exclude other members and to merge already existing groups in three separated stages for 15 rounds. An additional 15-period segment with new player IDs is then played. The

[10] Masclet [2003] finds similar results. In his public goods experiment, subjects exclude their peers for two reasons. Subjects are willing to punish unfair behaviors and expect behavioral changes in response to exclusions.

[11] The authors control for the order of implementation by switching the order of treatments. Non-parametric tests show that the order of implementation matters in contribution levels, apart from the last periods. Participants received information about the structure of the current treatment only.

[12] Average contributions increase to around 85% of the initial endowment when ostracism is implemented as the first treatment, and around to 80% when second.

[13] Indeed, authors show that subjects cast votes mainly at the beginning and at the end of each treatment so as to tame defectors.

experimental design considers three treatments. In the "Main" treatment, the MPCR is decreasing over group size but the total social benefit of contributing (i.e., MPCR times group size) is increasing, such that the social optimum is for all nine members to form a single group. The "Capped Efficiency" treatment is identical to Main with the only difference that the social benefit of contributions is capped after the group size is equal or higher than four, such that there is no efficiency gain to form groups greater than size four. In the Baseline treatment, subjects are randomly assigned to fixed groups (3-person, 6-person and 9-person groups) and no regrouping is possible.

They find that contribution levels over the 15 rounds are higher in "Main" and "Capped Efficiency" than in the control treatment. In the Main treatment, subjects manage to create grand coalitions, while under "Capped Efficiency" group size is on average four, consistent with the cap on groups' return. Group dynamics tend to be more stable in the "Main treatment" than in the "Capped Efficiency" treatment: subjects are more likely to exit and expel other members in the latter than in the former treatment. The novelty of this experiment is the presence of redemption. Redemption gives the possibility for individuals who have made early mistakes to later join successful groups and to become highly productive members.

In general, the group size is strongly affected by the presence of ostracism (see also Sääksvuori [2014]). Whenever ostracism is allowed, the group size is larger and more stable over time than without ostracism. However, the expulsion of group members may be costly. Another way to prevent free-riders from being part of the group without incurring the cost of expelling them is to not let them enter the group in the first place. Ahn et al. [2008] and Ahn et al. [2009] compare restricted entry when exit is free to an environment with free entry when exit is restricted. The game consists in a 20-period non-linear public goods game, with a dominant strategy of contributing 3. Each period begins with subjects being asked if they wish to change groups before making any decision regarding their contributions for the provision of the public good. They are informed about the aggregate contribution at the group level in the previous 5 rounds, and the number of subjects in each group who had chosen to remain in the group from the previous period. The between-subjects experimental design includes three treatments that differ in entry and exit rules.

In the "Free Entry/Exit" treatment, subjects are free to migrate among groups without any restriction, as in Ehrhart and Keser [1999] and Robbett [2016]. In the "Restricted Entry" treatment, exit is unrestricted but entering a new group is conditioned to the majority approval of its members. On the other hand, in the "Restricted Exit" treatment,

entering is free, but exiting a group is conditional on the approval of group members. In both cases, those voting can see the past five contributions of the subject attempting to enter or exit the group.

In the first study Ahn et al. [2008], the public good is pure and subjects therefore have an incentive to form the grand coalition with all twelve participants in a single group. They find that the "Restricted Entry" treatment has the highest levels of contributions and the group size is smaller than in the other treatments. The restriction on entry decisively teaches applicants to increase their contributions. Once a low-contributor is rejected by members of another group, he/she would raise her contributions until he/she gets accepted. A restriction on exit instead is detrimental for cooperation as high-contributors are prevented from exiting groups of defectors. In this case, cooperation unravels from the early rounds as subjects denied exit retaliate, lowering their contribution levels. The higher one's past contributions, the greater the likelihood of being accepted. The opposite holds when the subject is attempting to exit a group.

However, since the restricted entry mechanism promotes small groups, their members do not benefit from the positive externalities that come from being in a large group with a pure public good. Thus, despite being successful in promoting cooperation, the restricted entry mechanism also had the effect of reducing the average earnings of the subjects in comparison to the Free Entry/Exit and the Restricted Exit conditions. In a follow-up study, Ahn et al. [2009] incorporate congestion and thus reduce the incentive to form large groups. They find that Restricted Entry raises average contributions and total earnings, eliminates the earnings disadvantage of contributors in comparison to free-riders, and reduces the level of congestion within the groups in comparison to the Restricted Exit and Free Entry/Exit conditions (where congestion is defined as surplus members in a group beyond the optimal membership level, given current contributions).

To summarize, much of the experimental literature on group formation has focused on testing signals that can maintain a strict separation between free-riders and cooperators. The result is twofold: homogeneous groups composed of cooperators sustain high contributions to the group account, while the others almost never cooperate. A different strand of the literature on group formation created the conditions for groups to remain heterogeneous and equipped cooperators with a technology to discipline free-riders. With monetary punishment, ostracism or entry restrictions, the contribution decisions of individuals who initially behaved as free-riders change over time to converge to the decisions made by the cooperative subjects. This literature also highlights the conditions under

which punishment and exit restrictions have a dark side.

A closely related strand of the endogenous group formation literature considers situations in which participants can exit a specific relationship or avoid the game entirely. In such scenarios, agents may be able to opt-out of the social dilemma permanently (instead accepting payoffs that do not depend on the actions of others), "sit out" temporarily and then re-enter to continue interacting with the same partner(s), or exit the relationship and be re-matched with new partners.

2.3 Endogenous sorting with the option to play the game

The earliest experiment that we are aware of concerning this type of voluntary interaction is reported in Orbell et al. [1984].[14] They consider an n-player prisoner's dilemma game in which each of nine players chose whether to cooperate or defect, with each individual's payoffs strictly increasing in the number of cooperators and, conditional on the others' behavior, "defect" paying substantially more than "cooperate." After making their decision, each player chose whether to be paid based on the outcome of the game or to receive a stochastic payoff that did not depend on the behavior of the participants. They hypothesized that cooperators would be more inclined to exit than defectors, given that the expected outside option payoff is relatively more attractive to those choosing "cooperate" (who always earn lower payoffs from staying than defectors). In contrast, they find that those choosing the cooperative option were weakly less likely to exit. Though they find that cooperators have higher expectations regarding the cooperation of others, which could make staying more attractive, there is no evidence that this explains the difference in exit and they instead conclude that cooperators are simply more willing to be part of a group.

Orbell and Dawes [1993] consider one-shot, two-player prisoner's dilemma games and directly compare games in which participation is required to games in which participants could opt-out for a zero payoff – which was more attractive than the mutual defection outcome. They find that participants earned significantly more when participation was voluntary. Average cooperation was higher, conditional on playing the game, in the voluntary condition, as defectors tended to assume that their partners would defect as well and

[14] A precursor to this experiment is Miller and Holmes [1975], which introduced an "Expanded Prisoner's Dilemma" game that includes a third possible action for the players. This additional "defensive" or "withdrawal" option is a best response to one's partner defecting, but does not also reduce a cooperative partner's payoff. Although not a pure exit option, since payoffs to those electing to withdraw vary slightly depending on whether one's partner also withdrew, this expansion provides subjects with an option beyond simple cooperation or defection. They found that would-be cooperators are more likely to choose the withdrawal option than revert to defecting when facing an uncooperative partner. This experiment was not conducted under standard experimental economics procedures (participants actually played against a computer program rather than an actual partner and do not appear to have been incentivized).

thus were more likely than cooperators to choose not to play.

While these games were one-shot, Hauk [2003] considers a 10-period repeated prisoner's dilemma game in which participants can choose in each period whether to exit, defect, or cooperate. This experiment replicates the Orbell and Dawes [1993] finding that cooperation is higher when association is voluntary; the difference is that participants have knowledge of their prospective partners' history, and thus exit can be used to avoid future interactions with partners who have previously defected.

In the experiments discussed thus far, the payoff from exiting is typically higher than that from full defection, making universal non-entry and defection the subgame perfect equilibrium. Therefore, anyone who does enter the game, either with cooperative or competitive intentions, must do so expecting some degree of cooperation: an expectation consistently shown to be more prevalent among cooperators. When the outside option is worse than mutual defection, however, everyone will prefer to enter and thus exit provides a severe punishment for an uncooperative partner. This is explored in more detail in a third treatment reported by Hauk [2003], in which the mutual defection outcome is more attractive than not playing the game. While she finds evidence that subjects use the unattractive outside option as punishment for defection, cooperation is far lower than when participation was mandatory or the outside option was attractive, likely in part due to the fact that payoffs from (Defect, Defect) were positive in this treatment and negative in the others.

Hauk and Nagel [2001] further expand the investigation of exit with an attractive outside option. They consider scenarios in which participation does not require mutual agreement by both partners but, rather, one player can unilaterally force the game to be played – a setup that supplies data on how those who prefer to exit will play the game when required. They find that defection rates are lowest when participants can opt-out, such that, conditional on being matched, cooperation rates are high. At the same time, however, participants who are involuntarily forced to play the game frequently end up cooperating eventually. As a result, they suggest that unilateral matching is most favorable for cooperation rates: mutual defectors opt-out while many of those who are forced to play by their partner eventually learn to cooperate and can be "reformed."

Keser and Montmarquette [2011] consider a chosen-effort team production experiment with convex effort costs, in which participants are either required to participate in teams of two (where they face a social dilemma) or are paid only based on their own effort unless both players agree to form a team for the period. As in the "attractive outside

option" experiments described above, participants should not voluntarily form teams if they anticipate players choosing their dominant strategy effort in the team remuneration setting – but opting into team production could be profitable if players are sufficiently cooperative. They find that players regularly opt-in to the team remuneration arrangement and, after voluntarily joining a team, exert effort close to the Pareto optimal levels – which is significantly greater than effort in the mandatory team treatment.

Boun My and Chalvignac [2010] conduct a five-person 20-period public goods game, in which participants can (temporarily) opt-out of the group in each period and instead receive an outside option payoff that is either just slightly above the per-person endowment or somewhat higher. They find little effect of voluntary participation on average contributions. However, they find some evidence that the decay in contributions over time is mitigated in the condition with a higher outside option, and they attribute this finding to a pattern in which higher contributors opt-out after observing free-riding, causing lower contributors to increase the contributions to attract the cooperators back to the group.

Nosenzo and Tufano [2017] study a two person, one-shot public goods game, in which they vary whether voluntary participation can influence contributions either through assortative matching or through threat of exit. They compare a baseline treatment, in which participation is mandatory, with treatments where participation occurs only by mutual agreement. In the unconditional treatment, participants first decide whether to participate or take an outside option payoff that is just slightly above the mutual free-riding payoff. In the conditional treatment, both players first decide on contributions and then decide whether to continue with the partnership (and receive the resulting payoffs) or to withdraw such that both are instead paid the outside option. They find a strong effect of the conditional participation treatment, with contributions more than doubling (even including those who don't participate) relative to the baseline. However, there is no difference between contributions in the baseline and unconditional treatments. To assess whether sorting was limited by the relatively low outside option – which caused most players to opt in – they conducted an additional robustness check treatment with a higher outside option, but still find no difference in contributions compared to baseline. Finally, they also use a sequential prisoner's dilemma game to elicit beliefs and cooperative preferences, and do observe a false consensus effect (with cooperators believing cooperation is more likely), but these beliefs do not seem to influence the choice to opt-in to the unconditional public goods game.

With respect to use of exit as a threat or punishment, Wilson and Wu [2017] study

an infinitely repeated joint production task with imperfect monitoring, varying whether players can unilaterally and permanently terminate the game and the value of the outside option. They observe higher cooperation when termination is possible, regardless of the outside option value.

Yamagishi [1988] and Herbst et al. [2015] each investigate voluntary participation in a team work task and both find that stronger performers prefer to work as individuals rather than be compensated based on the combined efforts of a team. In the experiment reported in Yamagishi [1988], teams of three participated in a real effort (decoding) task for which they were paid an equal share of the team's output. Prior to each work period, team members could exit the group and instead be paid based on their own performance. As expected, the most productive team members were more likely to exit the group compensation scheme, even when there was a cost to doing so. Likewise, Herbst et al. [2015] find that participants who exert high effort in a chosen effort Tullock contest, either due to intrinsic competitiveness or lower induced effort costs, are less likely to join an alliance with another player. Despite this, they find that participants who voluntarily join alliances supply greater effort than those who were exogenously assigned to participate in an alliance, indicating a positive effect of opting-in to a team.

While the games discussed thus far only allowed the participant to avoid interaction with a fixed partner – but not acquire a new one – a theoretical literature on prisoner's dilemma simulations (beginning with Schuessler [1989]) allows agents who sever their interactions with one partner to randomly select a different partner from the population. In an experiment that provides a link between this theoretical work on random re-matching and the experiments on exit discussed thus far, Barclay and Raihani [2015] study a prisoner's dilemma game with costly punishment, in which participants could either: (1) withdraw from the game for a single period (receiving a payoff below the mutual defection payoff) and then return to the same partner; (2) withdraw for a single period and then be randomly matched with a new partner upon their return; or (3) pay a cost to be immediately matched with a new partner. They find that cooperation is lower when the participant can only withdraw and not switch to a new partner, in which case they are more likely to respond to a defection by punishing than using the withdrawal option. When changing partners is possible, participants are equally likely to switch partners or punish their existing partner.[15]

[15] Boone and Macy (1998) report an experiment in which subjects play a prisoner's dilemma-like online card game against simulated partners, with some players given the option to exit a relationship and be randomly paired with a new partner at any time. (Participants were not paid their payoffs, but, rather, the top performers received a fixed cash prize.) They do not find a difference in cooperation compared to

To summarize, we have seen evidence that endogenous sorting with the option to play the game, or the ability to exit, often promotes cooperation in social dilemmas among those who opt-in to the game or partnership. In games where participants can opt not to play (either at all or temporarily), the outside option is typically attractive relative to the Nash outcome of the social dilemma. Therefore, there is a selection effect: cooperators have higher expectations about the likelihood that others will cooperate and are more likely to opt-in to a social dilemma. The experiments reported by Orbell and Dawes [1993], Hauk [2003], Hauk and Nagel [2001], and Keser and Montmarquette [2011] all exhibit evidence of this type of selection or "forward-induction" argument. The downside is that those who opt-out do not learn to cooperate. There is additional indication that the positive effect of an exit option depends on the relative attractiveness of the outside option, although outside options that are less attractive than the unique Nash equilibrium have not yet been thoroughly studied. In cases where participants differ in their abilities or costs, there is considerable evidence that weaker players are more likely to opt-in to team remuneration schemes (Yamagishi [1988]; Herbst et al. [2015]; Keser and Montmarquette [2011] also find indication of this, but the result is not significant at the 10% level; Hamilton et al. [2003] observe a similar phenomenon in a field study of a garment factory). Finally, we also observe preliminary evidence that voluntary association in itself can increase team effort Herbst et al. [2015] and that the option to exit could increase cooperation further if exiting agents have the opportunity to be re-matched with a new partner, at least when punishment is also available Barclay and Raihani [2015].

3 Exogenous sorting

The endogenous sorting literature prompts the question of whether there is an intrinsic value to the freedom to choose one's partners.[16] After all, the separation of cooperators from free-riders could also be done exogenously, e.g. by the experimenter. The exogenous sorting of participants into groups has been the focus of much experimental research over the last decade.

Gunnthorsdottir et al. [2007] experimentally investigate the way individual disposition and history (experience playing with subjects having different preferences) interact to affect

the no-exit condition, and in a follow-up study they propose that this is due to exit having an opposing effect on the behavior of cooperative and uncooperative individuals: competitive players take advantage of the ability to escape retribution and use a "hit-and-run" strategy, while cooperators use the re-matching to form stable relationships with one another Boone and Macy [1999].

[16] There is a large literature on the value of free-choice in public goods games that is summarized in DeAngelo et al. [Forthcoming].

a subject's cooperative decision-making. The authors implement a public goods game with groups of 4 playing for 10 rounds. Each session includes 12 subjects. Their experiment has two main design features. First, subjects are grouped according to two different rules: (1) in the *baseline* condition, in each round, each subject has an equal chance of being grouped with any three other subjects, while in the (2) *sorted* condition, after subjects have made their contribution decisions, the four highest contributors are placed into one group, the fifth to the eighth highest are placed into another group and the four lowest into a third group. To avoid strategic behavior, subjects are not informed about the sorting mechanism. The second design feature of their experiment is that the two group assignment rules are crossed with three MPCR levels: 0.3, 0.5, and 0.75.

The authors find that within each MPCR condition, aggregate contribution levels in the sorted condition always exceed those in the baseline. Moreover, cooperative decay is slower in the sorted condition. To better understand the effect of history on individual decision-making in the sorted condition compared to the baseline, subjects were classified into *free-riders* if they contributed 30% or less of their endowment to the public account in the first round,[17] and into *cooperators* otherwise. An analysis by types reveals that cooperator contributions in the sorted condition exceed cooperator contributions in the baseline no later than the fourth round, and continue to do so until round 10. Accordingly, the authors conclude that cooperators behave differently in the sorted condition than in the baseline condition because in the former there are less encounters with free-riders than in the latter.

Gunnthorsdottir et al. [2010] inform the participants of the sorting mechanism, which generates an equilibrium in which all but two or three of the twelve participants in the session contribute fully (with the others contributing nothing). They find that contributions tend to stabilize around the efficient equilibrium level and are sustained for the duration of the 80 round session. Nax et al. [2017] generalize Gunnthorsdottir et al. [2010]'s design to include noise in how precisely each individual's contribution is detected. In doing so, they vary the degree to which group assignment accurately reflects earlier contributions, ranging from a situation in which groups are not based on earlier contributions (reflecting a "perfect strangers" public goods game) to the "perfect meritocracy" situation of Gunnthorsdottir et al. [2010]'s sorted condition. With no or low levels of noise, consistent with Gunnthors-

[17] Although the subjects were informed that the experiment had 10 rounds, the authors contend that first-round decisions reveal subjects' predisposition to cooperate because they are not influenced by the history of the game. However, as noted by Kreps and Wilson [1982], forward looking subjects may behave strategically starting from the first round with the intention to influence others' beliefs about the composition of the group and their subsequent behavior in the game.

dottir et al. [2010], contributions are close to the social optimum. Furthermore, even with substantial noise, such that the unique Nash equilibrium is full free-riding, contributions tend to stabilize at intermediate levels.

Gunnthorsdottir et al. [2007]'s approach is carried further by Ones and Putterman [2007], who control group formation in a more complex collective action environment. In their experiment, subjects have two decision variables under their control: (1) decisions on how much to contribute to the public account and (2) decisions on whether and by how much to impose costly punishment on other group members after learning of their contributions. More specifically, in each session, 16 subjects make 25 sets of contribution and punishment decisions. The 25 rounds are divided into 4 segments. That is, subjects are informed that they would be put in one group of 4 in the first round, a possibly different group whose members would not change during rounds 2–5, another fixed group for rounds 6–15, and a final fixed group for rounds 16–25. Ones and Putterman [2007] look at two different grouping rules: purposeful and random grouping.

The purposeful grouping treatment attempted first to identify types. In the first round, subjects made contribution decisions and were informed about the decisions made by the other three members of the group. However, without subjects' knowledge, groups were formed only after the first contribution decisions. This was done in order to create similarly diverse groups composed of one of the four highest first-round contributors in the session, one of the four lowest contributors, and so on. The assignment to groups for rounds 2-5 is similar to the first round. In addition, the computer also calculates a punishment index based on how much subjects punished low contributors at the end of the first round. Thus, groups are composed of both high and low contributors, both aggressive punishers of low contributors and non-punishers or perverse punishers. At the end of the fifth round, the subjects were ranked by their average contribution over rounds 1–5 as a whole and by their average punishment index over periods 1–5 as a whole. The contribution and punishment ranks were added together. Subjects with the lowest summed rank (high contributors who aggressively punished low contributors) were grouped *together* for rounds 6–15. The group composition is therefore homogeneous during these rounds.[18] Finally, in rounds 16–25 subjects were placed in randomly formed groups. In the random grouping treatment, subjects were placed in randomly created groups in rounds 1, 2, 6, 16.

Ones and Putterman [2007] confirm the findings from Gunnthorsdottir et al. [2007] that in the purposeful grouping treatment highly cooperative groups contribute more than their

[18]The authors wanted to have diverse groups in rounds 1–5 in order to control for differences in behavior stemming from differences in the kind of group one may find oneself in.

counterparts in the random grouping treatment. The authors also find that the behaviors displayed in rounds 1–5 are somewhat predictive of behaviors in rounds 6–15 and in rounds 16–25. The persistence of types over time holds for contribution and punishment decisions.

Although the classification method in Ones and Putterman [2007] and in Gunnthorsdottir et al. [2007] seem to be validated experimentally by the observed differences in behavior between the sorted and the unsorted groups, the effect size is not large. One of the possible reasons may lie in their sorting method that rests on repeated interactions which may introduce all sorts of strategic behavior. For instance, Ockenfels and Weimann [1999] report an experiment in which they sort people into "cooperative" and "less cooperative" groups on the basis of observing the subjects' contributions over ten rounds of a repeated public goods game. They do not observe any effect of this sorting on cooperation levels.

Gächter and Thöni [2005] implement an experimental protocol aimed at overcoming the possible bias introduced by measuring agents' types within repeated strategic interactions. They use a one-shot linear public goods game as the measurement instrument for cooperative attitudes. More specifically, subjects first participate in a "ranking" experiment, which consists of playing a one-shot linear public goods game with an MPCR $= 0.6$ in randomly formed groups of 3. Subjects were informed that they would play the "ranking" experiment just once and that some other part of the experiment would follow, but were not given further details. This was done in order to ensure that subjects' decisions in the ranking experiment are not strategically biased and reflect people's cooperative attitudes. Additionally, they did not receive immediate feedback on the outcome of the "ranking" game.

Subjects then received the instructions for the second part of the experiment, called the "sorted" experiment, which consists of playing a ten periods linear public goods game with fixed membership. Before playing the "sorted" experiment, subjects were informed that they had been ranked according to their contribution to the public account in the first part of the experiment, i.e. in the ranking experiment. They were also publicly informed that the three highest contributors in the ranking experiment were put together in one group, the next three in the second group and so on to the three lowest contributors who form the last group. Subjects were then informed about their new group members' contributions in the ranking experiment. There is also a control treatment, called the "unsorted" experiment, where after playing the ranking experiment, the groups are formed randomly and each participant is informed about their new group mates' contributions

in the ranking experiment. The authors also combine the two treatments – sorted and unsorted – with the opportunity to punish group members at a cost.

In the sorted experiment, the subjects who were sorted in the TOP contributor groups invested on average 18.1 ECU in the ranking experiment, MIDDLE contributors invested 10.1 ECU and LOW contributors invested 0.8 ECU. The sorting mechanism implemented by Gächter and Thöni [2005] led to a substantial increase in cooperation. While average contributions were 9.5 ECU in the unsorted experiment, they amounted to 13.9 ECU in the sorted experiment. At the individual level, the authors find that the top third of contributors in the sorted experiment invested significantly more in the public account than the most cooperative third of subjects in the unsorted experiment. Even without any punishment opportunities, the difference in cooperation levels is quite substantial (18.4 vs 14.1 ECU).[19]

The same year as the publication of Gächter and Thöni [2005]'s paper, there appeared Burlando and Guala [2005]'s experiment that combines multiple sources of evidence in order to refine the categories of agents in social dilemma games. Their experiment consists of two parts, with exactly a one week interval between them. The first part is composed of four different tasks: the strategy method task used by Fischbacher et al. [2001] to elicit participants' full schedule of contributions conditional on the cooperation of others; the decomposed game technique used by Offerman et al. [1996] to elicit social value orientation; a linear repeated public goods game for 20 rounds; and a questionnaire. The second part of the experiment consists of another 20 rounds of a repeated linear public goods game.

In the first part of the experiment, the assignment of subjects to groups was random. At the end of this part and before playing the second part of the experiment a week later, subjects were not given any information about the results of the various tasks. They were also unaware of the fact that their behavior in the four different tasks would be used to classify them into "types" and that similar types would end up in the same group for the second part of the experiment.[20]

With the data from the four different tasks, Burlando and Guala [2005] classify the subjects into three main categories: 32% are classified as free riders, 18% as (unconditional) cooperators, and 35% as reciprocators, with the remaining 15% classified in a "noisy"

[19] In fact, the availability of costly punishment does not result in higher cooperation levels compared to the sorted experiment without punishment. In the latter, the groups manage to sustain the same level of cooperation as in the former.

[20] More precisely, Burlando and Guala [2005] assigned weights to the four classification methods. Given that the ultimate goal was to understand behavior in a repeated public goods game, they decided to put more weight on the public goods game data, according to the following formula: repeated public goods game 40%; strategy method 20%; decomposed game 20%; questionnaire 20%. When no classification reached the 50% level (and therefore in all cases of tie), they assigned the subjects to a "noisy" group.

group. By focusing on the behavior of pre-determined types in the second part of the experiment, the authors notice that free-riders start with a lower level of contribution, which tends to decay very quickly. Cooperators and reciprocators start with a very high level of contribution, which remains high throughout most of the game. The behavior of reciprocators is particularly impressive, with constant (and almost full) contributions until round 19.

However, it should be noted that the difference in cooperation levels between cooperators and reciprocators is small. As emphasized by the authors, this is probably due to the fact that they were not entirely successful in forming perfectly homogeneous groups. Given that the sorting method implemented by Burlando and Guala puts a high weight on people's behavior in the repeated public goods game, it may be that their groups of cooperators were "infiltrated" by other types. This is not the only difference between their design and Gächter and Thöni [2005]'s experiment. Contrary to the later, in Burlando and Guala [2005] the subjects are not aware of the group composition.

de Oliveira et al. [2015] study how providing information about the group composition affects the behavior of different social preference types. Their experiment consists of two waves of experimental sessions. In the first wave participants are invited via email to participate in an internet experiment. They play a one-shot linear public goods game with strategy method as in Fischbacher et al. [2001]. The subjects' decisions are used to classify them into two types: (1) "Selfish" and (2) "Conditional Cooperators."[21] The second wave of the experiment took place on a different day. Participants were invited to the laboratory and played a linear public goods game for 15 rounds in groups of 3. The laboratory experiment has two main design features. First, participants are grouped into either (i) homogeneous groups of all conditional cooperators (C, C, C), (ii) homogeneous groups of all selfish players (S, S, S), or (iii) heterogeneous groups of two of one type and one of another (S, S, C) and (C, C, S). The second design feature of de Oliveira et al. [2015] is that participants are explicitly told about the group composition prior to starting the experiment.[22] In the "No Information" treatment, participants are not informed about the group composition. In both treatments, participants see information about their own type. The composition of the group remains unchanged for the 15 rounds.

The first surprising result is that under no information about the group composition contributions appear similar in the (C, C, C) grouping compared with the (C, C, S), but

[21] Subjects who never give more than five are classified as selfish. The threshold is similar to the one used by Gunnthorsdottir et al. [2007] to identify the selfish types.

[22] To be more precise, they are first told the meaning of the different types, their own type and the type of their mates.

higher compared with the (C, S, S). Given the small size of the group, the similar levels of cooperation in groups of three and groups of two conditional cooperators is indeed a puzzling finding. A plausible explanation may be that the classification method adopted by de Oliveira et al. [2015] does not distinguish between reciprocators and unconditional cooperators. The latter type of players would not respond to observed free-riding by reducing their own contributions. Another explanation may be that in small groups of three, the free-rider would still contribute positive amounts for strategic reasons.

The second original finding is that information about the group composition affects people's behavior solely in the (C, C, C) grouping. This suggests that for conditional cooperators, knowing that the other group mates are of the same type increases their individual investment in the group account. In addition to the actual presence of subjects with pro-social preferences, the existence of *common knowledge* that all the group members are of the same social preference type increases cooperation among like-minded people.[23]

To summarize, we have seen that the exogenous formation of groups by the experimenter, either based on contributions in earlier rounds under complete information (as in Gunnthorsdottir et al. [2010] and Nax et al. [2017]) or without knowledge of the sorting mechanism (as in Gunnthorsdottir et al. [2007] and Ones and Putterman [2007]) or based on earlier tasks that participants did not know would influence later groupings (as in Gächter and Thöni [2005], Burlando and Guala [2005], and de Oliveira et al. [2015]), regularly increase and sustain contributions in public goods games, particularly among the most cooperative participants. It turns out that, among cooperators, cooperation is higher when there is common knowledge about the group composition. Additionally, the method of measuring people's cooperative attitudes (i.e., the dimension on which they are sorted) plays a fundamental role in how successful groups are in sustaining cooperation. Some methods that elicit cooperativeness in non-repeated interactions (e.g. Gächter and Thöni [2005]) achieve a higher degree of homogeneity than others (e.g. Ockenfels and Weimann [1999]). The type-composition of the generated groups seems to explain the difference in the effect size between studies that use one-shot public goods games or multiple methods to sort out subjects.

[23]Other experiments where subjects are sorted out exogenously include van den Berg et al. [2015] who find that when grouped together frequency-based learners cooperate more than success-based learners, Kimbrough and Vostroknutov [2015] and Kimbrough and Vostroknutov [2016] introduce an incentivized method of eliciting individual norm-sensitivity and show that when grouped together the subjects who suffer more disutility from violating norms also behave more prosocially wherever there is a norm of prosocial behavior, e.g. in public goods games and common pool resource games. Finally, Junikka et al. [2017] and Nagatsu et al. [2018] find contradictory results about the effect of assortment information on cooperation within the cooperative and non-cooperative individuals, while Chaudhuri et al. [2016] find that conditional cooperators increase their contributions when informed about the presence of other conditional cooperators in the group.

4 Meta-analysis of cooperation under endogenous and exogenous sorting

In this section, we use an empirical strategy to compare the effects of endogenous sorting and exogenous matching on cooperation and to explore whether the two sorting mechanisms lead to differences in the probability for cooperators to be matched with like-minded partners. To do this, we collected data from existing papers studying public goods contributions under endogenous and exogenous sorting. We first present the searching and selection criteria, followed by a general overview of the effect of the two sorting mechanisms on cooperation levels. Then, in both mechanisms, we estimate the behavioral types and the likelihood of cooperators being matched into majority-cooperative groups.

Searching and selection criteria, and data: We include papers that comply with the following criteria:

- Incentivized laboratory experiments in a controlled environment;
- Repeated linear public good games such that:
  - The game is symmetric (homogeneity in the endowment);
  - The treatments concern the mechanism of sorting subjects into groups (endogenous or exogenous mechanism);
  - The subgame perfect Nash equilibrium is unique and Pareto dominated;
  - The group size > 2;
  - Decisions are taken simultaneously.

The search of the economics literature for studies meeting the above criteria was undertaken through Google Scholar, Internet Documents in Economics Access Service (IDEAS), references cited in Ledyard [1995] and Chaudhuri [2011], and posting messages on the Economic Science Association (ESA) Google group. Once the list of all the eligible studies was completed, we sent emails to the respective authors asking for the raw data from the experiment. As results, we collected group- and individual-level data from 12 different papers (out of 14 that satisfy our selection criteria) involving 17 treatments overall across 77 sessions (see Table 2). The excluded 2 datasets were either not available or it was not possible to follow groups' dynamics with unique IDs. It is also worth noting that no study implementing an endogenous mechanism with outside option has been included, for this

category mainly considers 2-person Prisoner's Dilemma games. As the final outcome, we obtained a (unbalanced) group-panel dataset of 303 groups and a individual-panel dataset counting 1164 individuals.

For each of these studies we select the following variables, according to existing surveys (Ledyard [1995], Chaudhuri [2011]) and meta-studies on contribution decisions in linear public goods games (Croson and Marks [2000], Zelmer [2003] and Fiala and Suetens [2017]): the MPCR, group size, the sorting mechanism implemented, total number of rounds, initial endowment, presence of punishment and the amount of group and individual contributions in each round (table 1). We also collect information about feedback given to subjects prior or during the experiment. However, we do not include this set of variables in the analysis since papers considered are homogeneous with respect to the feedback provided to subjects (see table A1).

Table 1: Descriptive statistics of variables used in the efficiency analysis.

Variable	**Mean**	**Std. Dev.**	**Min**	**Max**
Efficiency	0.60	0.33	0	1
MPCR	0.49	0.19	0.13	0.8
Group Size	4.47	2.51	2	12
Endogenous	0.70	0.46	0	1
Punishment	0.25	0.43	0	1
	n=5141			

Table 2: List of the papers included in the meta-analysis.

Papers	**Treatments**	**Groups**	**Individuals**	**MPCR**	**Rounds**	**Mechanism**	**Punishment**
Brekke et al. [2011]	1	29	87	0.5	20	Endogenous	No
Burlando and Guala [2005]	1	17	68	0.5	20	Exogenous	No
Cabrera et al. [2013]	1	20	80	0.5	20	Endogenous	No
de Oliveira et al. [2015]	2	14	42	0.5	15	Exogenous	No
Gächter and Thöni [2005]	2	42	126	0.6	10	Exogenous	Yes
Gunnthorsdottir et al. [2007]	3	33	132	0.3/0.5/0.75	10	Exogenous	No
Gürerk et al. [2006]	1	14	84	*variable*	30	Endogenous	Yes
Gürerk [2013]	1	28	168	*variable*	30	Endogenous	Yes
Hauge et al. [2018]	1	39	117	0.5	20	Endogenous	No
Kimbrough and Vostroknutov [2016]	2	24	96	0.5	10	Exogenous	No
Nicklisch et al. [2016]	1	27	100	*variable*	32	Endogenous	Yes
Page et al. [2005]	1	16	64	0.4	20	Endogenous	No
Total	17	303	1164				

Note: MPCR varies according to the group size in most of the experiments implementing endogenous sorting mechanisms.
Additionally, Gunnthorsdottir et al. [2007] implements 3 treatments, varying the MPCR.
Data from Nicklisch et al. [2016] consider only the treatment under perfect information (see paper, treatment *ONE*).
Data from de Oliveira et al. [2015] consider only the treatments aiming at creating homogeneous groups (see paper, treatments *CCC, FFF*).
Data from Hauge et al. [2018] consider only the study 4 as it is the only that matches our criteria (see paper, table 1)

Results from the meta-regression: Table 1 reports the main descriptives of the sample. The variable of interest for our analysis is the group contribution as a percentage of its total initial endowment. We will refer to this variable as *efficiency* reached by each group.[24] Figure 1 presents the contributions over time under both mechanisms. First, we note that efficiency levels are somewhat higher than what is observed in typical linear public goods experiments without sorting, where contributions typically begin at 40-60% of the endowment and decline significantly over time. The group-level means across all rounds are around 60% in exogenously-formed groups and 56% in endogenously-formed groups. While contributions under both mechanisms start at similar levels (table A2),[25] the pattern of contributions in endogenously-formed groups appears more stable over time. Specifically, exogenous sorting groups experience a marginal decline in efficiency over time, while efficiency in endogenous sorting groups increases, if anything, slightly.[26]

To formally explore the efficiency achieved under both mechanisms, we exploit the information provided by the panel data. To account for the dependence of the data, we estimate a multi-level model with clusters at the level of the groups, session, and paper (as recommended by Moffatt [2015]).[27] The estimates are reported in Table 3.

Model (1) regresses *Efficiency* only on the dummy variable *Endogenous*, which indicates that mechanism implemented in the experiment is endogenous, rather than exogenous, sorting. This model finds that there is no significant difference in overall efficiency between groups that were formed endogenously or sorted exogenously by the experimenter ($p = 0.94$). The second model (Col. 2) includes as controls features of the group that have been consistently found to influence public good contributions. The dummy variable *Punishment* indicates whether sanctioning opportunities were possible in a given group, the variable *MPCR* represents the value of the marginal per-capita return from the public good, and the variable *Group Size* indicates the number of members of the group in the current period. Note that these three variables are defined at the group – rather than session or paper – level. For instance, punishment is equal to zero if the subject's current group does not have access to the punishment mechanism, even if punishment were available to other groups within the same session. Model (2) indicates that the coefficient on *Endogenous*

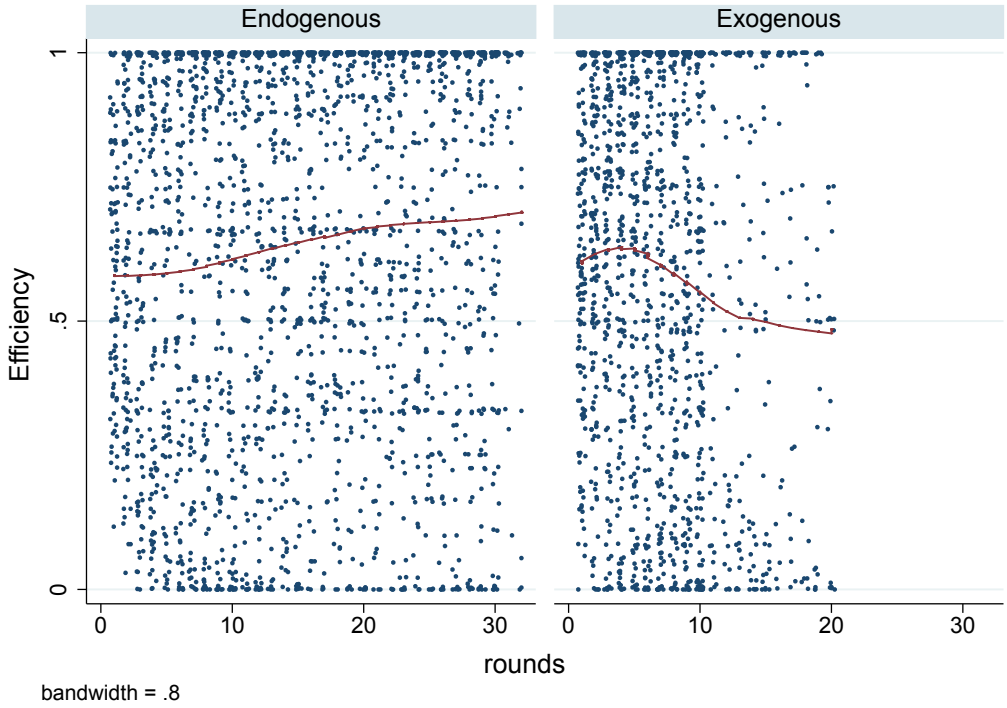
[24] Indeed, it can be seen as the relative distance to the highest achievable group outcome. Formally $\frac{Group\ Contribution}{Group\ Size \times Endowment}$.

[25] The difference in first-round contribution is not statistically significant ($t\text{-}statistic = -0.228$))

[26] While the Spearman correlation coefficients indicate the patterns over time are highly significant at the group-level ($\rho = -0.167$; $p < 0.001$ for exogenous and $= \rho = 0.20$; $p < 0.001$ for endogenous), when we cluster by paper we find a marginally significant decline in the exogenously-formed groups ($p = 0.065$) and no significant change in the endogenously-formed groups ($p = 0.155$).

[27] The results are quantitatively similar in a weighted least squares model with group level clustering (Table A3), as described in footnotes below.

Figure 1: LOWESS scatterplot of contributions over rounds.



Notes: Burlando and Guala [2005] is the only study with an exogenous sorting that lasts more than 15 rounds.

becomes somewhat more negative but remains statistically insignificant ($p = 0.35$) with these controls in place, while the coefficients on *Punishment*, *MPCR*, and *Group Size* are all positive and statistically significant, consistent with previous findings (Ledyard [1995], Nosenzo and Tufano [2017], Chaudhuri [2011]).

However, the first two models may mask a differential effect of punishment in endogenously and exogenously-formed groups. We hypothesize that punishment may be more effective at increasing efficiency in endogenously-formed groups. The reasons are two-fold. First, punishment and exogenous sorting may be substitutes, rather than complements, for increasing cooperation, and therefore punishment may play less of a role in groups that have already been sorted. This is consistent with the findings of Gächter and Thöni [2005], who find that punishment does not have a significant effect on contributions in either exogenously-formed groups of high cooperators (who cooperate regardless of punishment opportunities) or exogenously-formed groups of low cooperators. Second, consistent with the findings of Gürerk et al. [2006], cooperators who can sort endogenously may be more willing to opt-in to punishment environments, while free-riders avoid these situations.

To test this hypothesis, model (3) interacts the indicator for *Punishment* with *Endogenous*. The estimates show that punishment has little effect in experiments characterized

by exogenous sorting, but that there is a strong positive interaction between punishment and endogenous group formation. We also note the negative coefficient on Endogenous. While not reaching conventional levels of significance in the multi-level model (p=0.12), the estimate no longer resembles a precisely estimated zero as in the first model.[28] To determine whether this apparently negative effect of endogenous sorting in groups without punishment is due to free-riders opting into sanction-free groups, model (4) includes an additional indicator for *Sanction Free Groups*, which is 1 if the participant is in a group without punishment when groups with punishment were accessible in the session.[29] In this case, we observe a strong negative effect on efficiency in groups where subjects could avoid sanctions and the estimate on *Endogenous* is again close to zero and far from significant ($p = 0.98$). Finally, the interaction between Punishment and Endogenous remains significantly positive in this model, consistent with the positive effect of cooperators opting into sanctioning groups. We have therefore seen that punishment is more effective in endogenously-formed groups due to a selection effect, in which cooperators seek out punishment institutions and free-riders avoid them.[30]

[28] This estimate is significantly negative at the $p = 0.02$ level in the weighted least squares specification.

[29] The number of total groups belonging to this category amounts to 31, for a total of 4666 observations.

[30] Finally, we can also consider punishment as a session-level, rather than group-level, variable that indicates that punishment is available to the subjects but may or may not be a feature of their current groups (e.g., they could freely join a different group where punishment is possible). If we substitute this indicator for the group-level punishment indicator and its interaction with *Endogenous*, we find that the positive interaction disappears and, like exogenously-formed groups, punishment as a session-wide feature does not have a positive effect in endogenously-formed groups. We thus conclude that punishment is effective in groups with sorting only if it can be used as a further means of enabling cooperators to sort into groups free of free-riders.

Table 3: Multi-Level Model

	(1)	(2)	(3)	(4)
Endogenous	-0.004	-0.065	-0.142	-0.001
	(-0.08)	(-0.93)	(-1.52)	(-0.02)
Punishment		0.364***	0.067	0.084
		(8.49)	(1.00)	(1.30)
MPCR		0.390***	0.388***	0.381***
		(8.18)	(8.14)	(8.00)
Group Size		0.036***	0.035***	0.035***
		(10.15)	(9.94)	(9.88)
Endogenous*Punishment			0.500***	0.174*
			(5.92)	(1.86)
Sanction-Free Groups				-0.345***
				(-4.92)
Constant	0.574***	0.196***	0.230***	0.231***
	(14.37)	(3.03)	(2.88)	(4.28)
Observations	5141	5141	5141	5141

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Dependent variable is *Efficiency*. Multi-level model with group, session and paper clusters.

We further investigate one of the main reasons for the absence of any difference in cooperation between endogenously sorted groups and exogenously sorted ones: the likelihood of cooperators being matched into majority-cooperative groups. To do this, we first identify the intrinsic types of the subjects and then examine whether either of the sorting mechanisms favors positive assortment more than the other. In light of the efficiency analysis, we do not expect any difference between mechanisms in terms of the probability of matching like-minded partners.

In the public goods literature, it is common to classify subjects into 3 types: free-riders (FR), conditional cooperators (CC) and unconditional cooperators (UC) by implementing the strategy method (Fischbacher et al. [2001], Fischbacher and Gächter [2010]). Unfortunately, the vast majority of the considered studies do not elicit subjects' conditional contributions profile with the strategy method, and it is impossible to retrieve this information from the data collected. For this reason, we implement a technique based on first round contributions to distinguish subjects into cooperators – irrespective to their conditional nature – and free-riders (as in Gächter and Thöni [2005] and de Oliveira et al. [2015] among others).[31] We classified subjects as either cooperators or free-riders whether their first-round contribution is higher than a given threshold, namely 50% of the endowment.[32]

The resultant distribution of types is reported for each paper in Table 4. In papers implementing endogenous sorting, the algorithm classifies on average 31.4% of the subjects as free-riders, 68.6% as cooperators. Similarly, under exogenous sorting, the algorithm classifies on average 31.6% of the subjects as free-riders, and 68.4% as cooperators. Shares of cooperative and free-rider types do not differ between endogenous and exogenous sorting (Mann-Whitney rank-sum test on percentages, $p = 0.94$). Furthermore, shares of cooperative and defective types are in line with the related literature (Fischbacher et al. [2001], Fischbacher and Gächter [2010], Kurzban and Houser [2005]).[33]

We now aim at shedding light on groups' composition using types estimates. In par-

[31] Despite its large use in the experimental economics literature, we acknowledge the limits of this approach (see for example Burton-Chellew et al. [2016]) and in particular the fact that this technique may confound conditional cooperators with free-riders. Chaudhuri and Paichayontvijit [2006] show that often what appears as free-riding is also conditional behavior predicated on pessimistic prior beliefs about peer contributions. Conditional cooperators who are pessimists will make low contributions, which make them appear to be free-riders. We also acknowledge that subject behavior throughout the experiment may deviate from these initial type classifications based on first-round contributions, as free-riders may mimic cooperators or conditional cooperators may begin to free-ride after being exposed to uncooperative subjects.

[32] We choose this threshold since it clearly cuts the distribution of first-round contributions in two (see Figure B1). We also employed lower thresholds obtaining similar results in terms of type distribution.

[33] Additionally, there are no significant differences in the percentage of participants classified as cooperators based on whether punishment is available in their initial group or punishment is available in the session as a whole, and no significant interactions between punishment and sorting mechanism.

Table 4: Shares of estimated subjects types from the first-round contribution technique.

Papers	**% Free-Riders**	**% Cooperators**
Exogenous		
Burlando and Guala [2005]	19.2	80.8
de Oliveira et al. [2015]	52.4	47.6
Gächter and Thöni [2005]	25.4	74.6
Gunnthorsdottir et al. [2007]	34.1	65.9
Kimbrough and Vostroknutov [2016]	27.1	72.9
Average Exogenous	31.6	68.4
Endogenous		
Brekke et al. [2011]	25.3	74.7
Cabrera et al. [2013]	21.2	78.8
Gürerk et al. [2006]	43.9	56.1
Gürerk [2013]	43.7	56.3
Hauge et al. [2018]	27.3	72.7
Nicklisch et al. [2016]	47.8	52.2
Page et al. [2005]	10.9	89.1
Average Endogenous	31.4	68.6
Average Overall	31.5	68.5

ticular, we focus on the probability that a cooperative type joins a group mostly made of like-minded subjects.

In the following comparative analysis, groups are classified in each round as cooperative or uncooperative based on whether the majority of their members are, respectively, cooperators or free-riders. We are interested in the frequency of "positive matching" between cooperative types and cooperative groups. We therefore define the indicator *Matching* which takes value 1 when a subject classified as cooperative joins a cooperative group, and 0 otherwise. According to the results reported in the efficiency analysis, we expect that a cooperator is equally likely to join cooperator-majority groups under the two different mechanisms. Additionally, we ought to observe a higher probability of positive assortment among cooperative types when punishment technologies are present at the group level. Table 5 reports the results from a multi-level, linear probability model.[34] Model (1) finds no difference between sorting mechanisms ($p = 0.79$). Significance levels remain similar when controlling for other important factors affecting the matching of cooperative partners, such as the group size, punishment and the percentages of cooperative types in a given session (Models 2-5). Model (2) controls for the size of the group. As expected, it turns out that *Group Size* affects negatively the probability of matching cooperative subjects, and under this model specification the estimate associated to *Endogenous* is still

[34]Clustering at the paper, session, group and individual level. Results are quantitatively similar in a logit model (see table A4).

not significant ($p = 0.45$). In Model (3), we account for the slight variation in the number of cooperators among sessions, introducing the variable *% Cooperators*. A higher share of cooperators in the session increases the probability of successful matching, while we again do not find any difference between sorting mechanisms ($p = 0.12$). Model (4) instead controls for group-level punishment, distinguishing the effect across sorting mechanisms. The model finds no evidence of the effect of *Endogenous* ($p = 0.84$), *Punishment* ($p = 0.68$) and the interaction term *Endogenous∗Punishment* ($p = 0.14$). However, when including *Group Size* and *% Cooperators* (Model 5), consistent with the findings reported in the efficiency analysis, punishment facilitates the matching between like-minded cooperators solely under the endogenous sorting mechanism ($p = 0.06$).[35] With these variables in place, the parameter associated with *Endogenous* is far from being statistically significant ($p = 0.34$).

Results from the analysis above provide strong evidence showing that a cooperative subject has not extra chances to join majority-cooperator groups in either sorting mechanism. What is more, results confirm the beneficial effect of punishment on efficiency levels by facilitating the assortment of cooperators in the same groups.

Table 5: Linear probability multi-level regression

	(1)	(2)	(3)	(4)	(5)
Endogenous	0.027	0.062	0.067	-0.026	0.048
	(0.27)	(0.76)	(1.57)	(-0.20)	(0.95)
Group Size		-0.024***			-0.024***
		(-16.92)			(-17.08)
% Cooperators			1.264***		1.326***
			(9.94)		(9.74)
Punishment				0.045	0.055
				(0.41)	(0.64)
Endogenous*Punishment				0.200	0.203*
				(1.46)	(1.86)
Constant	0.695***	0.780***	-0.188*	0.689***	-0.151
	(9.11)	(12.13)	(-1.96)	(7.00)	(-1.46)
Observations	15095	15095	15095	15095	15095

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t statistics in parentheses. The dependent variable is the indicator *Matching*. Clustering at the paper, session, group and individual level.

[35] The effect, despite reaching conventional levels under the multi-level linear probability model, is more significant under the logit model.

Our analysis has focused on the success of cooperators successfully matching into groups predominantly comprised of like-minded individuals, mirroring the focus of the reviewed papers. However, we can also briefly consider the behavior and outcomes of the free-riders. We find that free-riders are also more likely to sort into cooperative groups (while remaining in the minority) solely when both punishment and endogenous sorting are present.[36] As we have seen from the efficiency analysis, however, the presence of these free-riders does not negatively affect overall group cooperation in endogenously-sorted groups with punishment. Is it the case that those who were classified as intrinsic free-riders adjust their actual behavior over the course of the experiment and mimic their more cooperative group members? Consistent with this hypothesis, we find that free-riders do contribute significantly more in endogenously-formed groups than in exogenously-sorted groups where punishment is available (multi-level model, see table A6). Indeed, the interaction between Endogenous and Punishment is consistent with free-riders contributing an additional 50% ($p < 0.01$) of their endowment in these groups and is significant at all conventional levels. Thus, we find evidence that endogenously-formed groups that implement punishment are not only particularly successful in enabling cooperators to match with like-minded individuals and maintain high contributions, but they also allow free-riders to reform their behavior and benefit from this cooperation.

[36]The interaction of Endogenous and Punishment is significant at the $p < 0.01$ level for free-riders in the mixed-level model with clusters at the level of the individual, groups, session, and paper (table A5).

5 Conclusion

There are several important results that emerge from our paper. First, the literature on endogenous sorting suggests that cooperation may only be prevented from unraveling by screening and separating defectors from conditionally cooperative individuals. The introduction of costly signals – for example, a higher group entry fee – or the availability of the option to not play the game often promote cooperation in social dilemmas among those who end up in cooperative groups or partnerships. The downside is that the separating equilibrium achieved in some experiments prevent free-riders to learn to cooperate. The introduction of monetary punishment, ostracism, and entry or exit restrictions allows cooperation to be sustained in mixed groups of cooperators and selfish subjects. The latter type are compelled to change their behavior when cooperators dispose of enforcement mechanisms.

The second important lesson that emerges from the surveyed literature is that contrary to the literature showing that endogenous choices encourage cooperation compared to when the same environment is exogenously imposed (e.g. Dal Bo et al. [2010]), we find no difference in terms of cooperation between studies implementing an endogenous sorting and experiments implementing an exogenous matching of subjects into groups. What is more, we found that it is no more likely for a cooperator to be matched with like-minded partners under endogenous sorting than under exogenous. The one significant difference we observe across mechanisms is that endogenous sorting better enables would-be free-riders to reform their behavior, contribute more, and co-exist alongside cooperators, particularly when supported by punishment. These observations are obviously related. As we showed in the survey sections, the success of a sorting method in matching like-minded individuals and the levels of cooperation are closely interlinked. The absence of differences in the assortment of cooperative individuals may be one of the main reasons for the absence of differences in cooperation levels between the two sorting mechanisms. However, it should be noted that most experiments with exogenous sorting included in our meta-analysis elicit people's cooperative attitudes in non-repeated games, which may result in more homogeneous groups compared to papers where subjects were sorted in groups on other dimensions (e.g., repeated games, psychological measures of personality traits).

References

T. K. Ahn, R. Mark Isaac, and Timothy C. Salmon. Endogenous group formation. *Journal of Public Economic Theory*, 10(2):171–194, 2008. doi: 10.1111/j.1467-9779.2008.00357.x.

T. K. Ahn, R. Mark Isaac, and Timothy C. Salmon. Coming and going: Experiments on endogenous group sizes for excludable public goods. *Journal of Public Economics*, 93 (1-2):336–351, 2009. doi: 10.1016/j.jpubeco.2008.06.007.

Jason A. Aimone, Laurence R. Iannaccone, Michael D. Makowsky, and Jared Rubin. Endogenous group formation via unproductive costs. *Review of Economic Studies*, 80(4): 1215–1236, 2013. doi: 10.1093/restud/rdt017.

Pat Barclay and Nichola Raihani. Partner choice versus punishment in human Prisoner's Dilemmas. *Evolution and Human Behavior*, 172:263–271, 2015. doi: 10.1016/j.evolhumbehav.2015.12.004.

Ralph Bayer. Cooperation in Partnerships: The Role of Breakups and Reputation. *Journalof Institutional and theoretical Economics*, 172:615–638, 2016. doi: 10.1628/093245616X14610627109836.

Iris Bohnet and Dorothea Kübler. Compensating the cooperators: Is sorting in the prisoner's dilemma possible. *Journal of Economic Behavior and Organization*, 56(1):61–76, 2005. doi: 10.1016/j.jebo.2003.04.002.

R. Thomas Boone and Michael W. Macy. Unlocking the Doors of the Prisoner's Dilemma: Dependence, Selectivity, and Cooperation. *Social Psychology Quarterly*, 62:32–52, 1999.

Kene Boun My and Benoît Chalvignac. Voluntary participation and cooperation in a collective-good game. *Journal of Economic Psychology*, 31(4):705–718, 2010. doi: 10.1016/j.joep.2010.05.003.

Kjell Arne Brekke, Karen Evelyn Hauge, Jo Thori Lind, and Karine Nyborg. Playing with the good guys. A public good game with endogenous group formation. *Journal of Public Economics*, 95(9-10):1111–1118, 2011. doi: 10.1016/j.jpubeco.2011.05.003.

Roberto M Burlando and Francesco Guala. Heterogeneous agents in public goods experiments. *Experimental Economics*, 8:35–54, 2005.

Maxwell N. Burton-Chellew, Claire El Mouden, and Stuart A. West. Conditional cooperation and confusion in public-goods experiments. *Proceedings of the National Academy of Sciences*, 2016. doi: 10.1073/pnas.1509740113.

Susana Cabrera, Enrique Fatas, Juan A Lacomba, and Tibor Neugebauer. Vertically splitting a firm: promotion and relegation in a team production experiment. *Experimental Economics*, 16:426–441, 2013.

Gary Charness and Chun Lei Yang. Starting small toward voluntary formation of efficient large groups in public goods provision. *Journal of Economic Behavior and Organization*, 102:119–132, 2014. doi: 10.1016/j.jebo.2014.03.005.

Ananish Chaudhuri. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, 14(1):47–83, 2011. doi: 10.1007/s10683-010-9257-1.

Ananish Chaudhuri and Tirnud Paichayontvijit. Conditional Cooperation and Voluntary Contributions to a Public Good. *Economics Bulletin*, (3):1–15, 2006.

Ananish Chaudhuri, Tirnud Paichayontvijit, and Alexander Smith. Belief heterogeneity and contributions decay among conditional cooperators in public goods games. *Journal of Economic Psychology*, (58):15–30, 2016. doi: 10.1016/j.joep.2016.11.004.

Matthias Cinyabuguma, Talbot Page, and Louis Putterman. Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics*, 89(8 SPEC. ISS.):1421–1435, 2005. doi: 10.1016/j.jpubeco.2004.05.011.

Giorgio Coricelli, Dietmar Fehr, and Gerlinde Fellner. Partner Selection in Public Goods Experiments. *Journal of Conflict Resolution*, 48(3):356–378, 2004. doi: 10.1177/0022002704264143.

Rachel T A Croson and Melanie Beth Marks. Step returns in threshold public goods: A meta- and experimental analysis. *Experimental Economics*, 2(3):239–259, 2000. doi: 10.1007/BF01669198.

Pedro Dal Bo, Andrew Foster, and Louis Putterman. Institutions and Behavior : Experimental Evidence on the Effects of Democracy. *American Economic Review*, 2010. ISSN 00028282. doi: 10.1257/aer.100.5.2205.

Angela C.M. de Oliveira, Rachel T.A. Croson, and Catherine Eckel. One bad apple? Heterogeneity and information in public good provision. *Experimental Economics*, 18 (1):116–135, 2015. doi: 10.1007/s10683-014-9412-1.

J. Gregory DeAngelo, Dimitri Dubois, and Rustam Romaniuc. The perils of democracy. *Journal of Economic Behavior and Organization*, Forthcoming.

Karl-Martin Ehrhart and Claudia Keser. Mobility and Cooperation: On the Run. *Working papers, Cirano*, 1999.

Lenka Fiala and Sigrid Suetens. Transparency and cooperation in repeated dilemma games: a meta study. *Experimental Economics*, 20(755):1–17, 2017. doi: 10.1007/s10683-017-9517-4.

Urs Fischbacher and Simon Gächter. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1):541–556, 2010. ISSN 00028282. doi: 10.1257/aer.100.1.541.

Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3):397–404, 2001. doi: 10.1016/S0165-1765(01)00394-9.

Simon Gächter and Christian Thöni. Social Learning and Voluntary Cooperation Among Like-Minded People. *Journal of the European Economic Association*, 3:303–314, 2005. doi: 10.1162/jeea.2005.3.2-3.303.

Veronika Grimm and Friederike Mengel. Cooperation in viscous populations-Experimental evidence. *Games and Economic Behavior*, 66(1):202–220, 2009. doi: 10.1016/j.geb.2008.05.005.

Anna Gunnthorsdottir, Daniel Houser, and Kevin McCabe. Disposition, history and contributions in public goods experiments. *Journal of Economic Behavior and Organization*, 62(2):304–315, 2007. doi: 10.1016/j.jebo.2005.03.008.

Anna Gunnthorsdottir, Roumen Vragov, Stefan Seifert, and Kevin McCabe. Near-efficient equilibria in contribution-based competitive grouping. *Journal of Public Economics*, 94 (11-12):987–994, 2010. doi: 10.1016/j.jpubeco.2010.07.004.

Özgür Gürerk. Social learning increases the acceptance and the efficiency of punishment institutions in social dilemmas. *Journal of Economic Psychology*, 34:229–239, 2013. doi: 10.1227/01.NEU.0000349921.14519.2A.

Özgür Gürerk, Bernd Irlenbusch, and Bettina Rockenbach. The competitive advantage of sanctioning institutions. *Science*, 312:108–111, 2006.

Özgür Gürerk, Bernd Irlenbusch, and Bettina Rockenbach. Voting with feet: community choice in social dilemmas. *Uni Erfurt Working Paper*, (4643):1–46, 2010.

Özgür Gürerk, Bernd Irlenbusch, and Bettina Rockenbach. On cooperation in open communities. *Journal of Public Economics*, 120:220–230, 2014. doi: 10.1016/j.jpubeco.2014.10.001.

Barton H. Hamilton, Jack A. Nickerson, and Hideo Owan. Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation. *Journal of Political Economy*, 111(3):465–497, 2003. doi: 10.1086/374182.

Karen Evelyn Hauge, Kjell Arne Brekke, Karine Nyborg, and Jo Thori Lind. Sustaining cooperation through self-sorting: The good, the bad, and the conditional. *Proceedings of the National Academy of Sciences*, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1802875115.

Esther Hauk. Multiple prisoner's dilemma games with(out) an outside option: An experimental study. *Theory and Decision*, 54(3):207–229, 2003. doi: 10.1023/A:1027385819400.

Esther Hauk and Rosemarie Nagel. Choice of Partners in Multiple Two-Person Prisoner's Dilemma Games: An Experimental Study. *Journal of Conflict Resolution*, 45(6):770–793, 2001. doi: 10.1177/0022002701045006004.

Luisa Herbst, Kai A. Konrad, and Florian Morath. Endogenous group formation in experimental contests. *European Economic Review*, 74:163–189, 2015. doi: 10.1016/j.euroecorev.2014.12.001.

Laurence R. Iannaccone. Sacrifice and Stigma: Reducing Free-riding in Cults, Communes, and Other Collectives. *Journal of Political Economy*, 100(2):271–291, 1992. doi: 10.1086/261818.

Jaakko Junikka, Lucas Molleman, Pieter Van Den Berg, Franz J. Weissing, and Mikael Puurtinen. Assortment, but not knowledge of assortment, affects cooperation and individual success in human groups. *PLoS ONE*, 2017. doi: 10.1371/journal.pone.0185859.

Claudia Keser and Claude Montmarquette. Voluntary versus Enforced Team Effort. *Games*, 2(3):277–301, 2011. doi: 10.3390/g2030277.

Claudia Keser and Frans van Winden. Conditional Cooperation and Voluntary Contributions to Public Goods. *The Scandinavian Journal of Economics*, 102(1):23–39, 2000. doi: 10.1111/1467-9442.00182.

Erik O. Kimbrough and Alexander Vostroknutov. The social and ecological determinants of common pool resource sustainability. *Journal of Environmental Economics and Management*, 72(430):38–53, 2015. doi: 10.1016/j.jeem.2015.04.004.

Erik O. Kimbrough and Alexander Vostroknutov. Norms Make Preferences Social. *Journal of the European Economic Association*, 14(3):608–638, 2016. doi: 10.1111/jeea.12152.

David M. Kreps and Robert Wilson. Sequential Equilibria. *Econometrica*, 50(4):863—-894, 1982. doi: 10.2307/1912767.

R. Kurzban and D. Houser. Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences*, 102(5):1803–1807, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0408759102.

John O Ledyard. Public Goods: A Survey of Experimental Research. In *The Handbook of Experimental Economics*, pages 111–194. 1995. doi: 10.3987/Contents-12-85-7.

Frank P. Maier-Rigaud, Peter Martinsson, and Gianandrea Staffiero. Ostracism and the provision of a public good: experimental evidence. *Journal of Economic Behavior & Organization*, 73(3):387–395, 2010. doi: 10.1016/j.jebo.2009.11.001.

David Masclet. Ostracism in work teams: a public good experiment. *International Journal of Manpower*, 24(7):867–887, 2003. doi: 10.1108/01437720310502177.

D T Miller and J G Holmes. The role of situational restrictiveness on self-fulfilling prophecies: A theoretical and empirical extension of Kelley and Stahelski's triangle hypothesis. *Journal of Personality and Social Psychology*, 31:661–673, 1975. doi: 10.1037/h0077081.

Peter Moffatt. *Experimetrics: Econometrics for Experimental Economics.* MacMillan International Higher Education, 2015.

Michiru Nagatsu, Karen Larsen, Mia Karabegovic, Marcell Székely, Dan Mønster, and John Michael. Making good cider out of bad apples Signaling expectations boosts cooperation among would-be free riders. *Judgment and Decision Making*, 13(1), 2018.

Heinrich H Nax, Stefano Balietti, Ryan O Murphy, and Dirk Helbing. A noisy institution : An experimental welfare investigation of the contribution-based grouping mechanism. *Social Choice and Welfare*, 2017. doi: 10.1007/s00355-017-1081-5.

Andreas Nicklisch, Kristoffel Grechenig, and Christian Thöni. Information-sensitive Leviathans. *Journal of Public Economics*, (144):1–13, 2016. doi: http://dx.doi.org/10.1016/j.jpubeco.2016.09.008.

Daniele Nosenzo and Fabio Tufano. The Effect of Voluntary Participation on Cooperation. *Journal of Economic Behavior & Organization*, 142:307–319, 2017. doi: 10.1016/j.jebo.2017.07.009.

Axel Ockenfels and Joachim Weimann. Types and patterns: an experimental East-West-German comparison of cooperation and solidarity. *Journal of Public Economics*, 71(2):275–287, 1999. doi: 10.1016/S0047-2727(98)00072-3.

Theo Offerman, Joep Sonnemans, and Arthur Schram. Value Orientations, Expectations and Voluntary Contributions in Public Goods. *The Economic Journal*, 106(437):817, 1996. doi: 10.2307/2235360.

Umut Ones and Louis Putterman. The ecology of collective action: A public goods and sanctions experiment with controlled group formation. *Journal of Economic Behavior and Organization*, 62(4):495–521, 2007. doi: 10.1016/j.jebo.2005.04.018.

John M Orbell and Robyn M. Dawes. Social Welfare, Cooperators' Advantage, and the Option of Not Playing the Game. *American Sociological Review*, 58(6):787–800, 1993. doi: 10.2307/2095951.

John M Orbell, P Schwartz-Shea, and Randy T Simmons. Do Cooperators Exit More Readily Than Defectors? *American Political Science Review*, 76(1):753–766, 1984. doi: 10.2307/1961254.

Talbot Page, Louis Putterman, and Bulent Unel. Voluntary Association in Public Goods Experiments :. *Economic Journal*, 506:1032–1053, 2005.

David G. Rand and Martin A. Nowak. Human cooperation. *Trends in Cognitive Sciences*, 17(8), 2013. doi: 10.1016/j.tics.2013.06.003.

Andrea Robbett. Community dynamics in the lab. *Social Choice and Welfare*, 46(3):543–568, 2016. doi: 10.1007/s00355-015-0928-x.

Lauri Sääksvuori. Intergroup conflict, ostracism, and the evolution of cooperation under free migration. *Behavioral Ecology and Sociobiology*, 68(8):1311–1319, 2014. doi: 10.1007/s00265-014-1741-8.

Rudolf Schuessler. Exit Threats and Cooperation under Anonymity. *The Journal of Conflict Resolution*, 33(4):728–749, 1989. doi: 10.1177/0022002789033004007.

Matthias Sutter, Stefan Haigner, and Martin G. Kocher. Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Review of Economic Studies*, 2010. doi: 10.1111/j.1467-937X.2010.00608.x.

Charles M. Tiebout. A Pure Theory of Local Expenditures. *Journal of Political Economy*, 64(5):416–424, 1956. doi: 10.1086/257839.

Pieter van den Berg, Lucas Molleman, and Franz J. Weissing. Focus on the success of others leads to selfish behavior. *Proceedings of the National Academy of Sciences*, 112 (9):2912–2917, 2015. doi: 10.1073/pnas.1417203112.

Alistair J. Wilson and Hong Wu. At-will relationships: How an option to walk away affects cooperation and efficiency. *Games and Economic Behavior*, 102:487–507, 2017. doi: 10.1016/j.geb.2017.02.007.

Toshio Yamagishi. Exit from the group as an individualistic solution to the free rider problem in the United States and Japan. *Journal of Experimental Social Psychology*, 24 (6):530–542, 1988. doi: 10.1016/0022-1031(88)90051-0.

Jennifer Zelmer. Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6(3):299–310, 2003. doi: 10.1023/A:1026277420119.

A Supplementary tables

Papers	Individual		Average		Ranking other groups	Non-contribution	
	Own group	Other groups	Own group	Other groups		Groups	Individuals
Brekke et al. [2011]	No	No	Yes	Yes	No	Yes	No
Burlando and Guala [2005]	No	No	Yes	No	No	No	No
Cabrera et al. [2013]	No	No	Yes	No	No	No	No
de Oliveira et al. [2015]	No	No	Yes	No	No	No	Yes
Gächter and Thöni [2005]	No	No	Yes	No	Yes	No	No
Gunnthorsdottir et al. [2007]	No	Yes	No	No	No	No	No
Gürerk et al. [2006]	Yes	No	Yes	Yes	No	No	No
Gürerk [2013]	Yes	No	Yes	Yes	No	No	No
Hauge et al. [2018]	No	No	Yes	Yes	No	Yes	No
Kimbrough and Vostroknutov [2016]	No	No	Yes	No	No	No	No
Nicklisch et al. [2016]	Yes	No	Yes	Yes	No	No	No
Page et al. [2005]	Yes	No	Yes	Yes	No	No	No

Table A1: Information given to subjects prior or during the experiment. *Individual* reports whether subjects learn about individual contributions in their own group or others. The same applies to *Average*, with the only difference of referring to average contributions. *Ranking other groups* asks if subjects learned an ordinal scale of groups based on their contributions. The category *Non-contribution* refers to characteristics of either groups or individuals that are not contribution related. We asked if subjects learned about some characteristic of the other group that is not contribution-related, and whether they learned about the type of the other group members (without necessarily knowing their exact contributions)

$Round$	$Efficiency_{Endo}$	$Efficiency_{Exo}$
1	0.610	0.603
2	0.624	0.631
3	0.577	0.648
4	0.581	0.644
5	0.573	0.638
6	0.564	0.638
7	0.582	0.609
8	0.572	0.615
9	0.566	0.591
10	0.546	0.519
11	0.612	0.485
12	0.568	0.474
13	0.603	0.467
14	0.594	0.447
15	0.575	0.464
16	0.606	0.634
17	0.597	0.554
18	0.560	0.540
19	0.564	0.468
20	0.519	0.399
21	0.672	-
22	0.701	-
23	0.657	-
24	0.673	-
25	0.664	-
26	0.697	-
27	0.697	-
28	0.673	-
29	0.687	-
30	0.671	-
31	0.833	-
32	0.701	-

Table A2: Average efficiency by round and mechanism

Table A3: WLS regression

	(1)	(2)	(3)	(4)
Endogenous	0.0565	-0.160**	-0.247***	-0.110
	(0.56)	(-2.33)	(-3.07)	(-1.60)
Punishment		0.308***	0.100	0.107
		(5.11)	(1.09)	(1.17)
MPCR		1.291***	1.298***	1.241***
		(4.60)	(4.52)	(4.17)
Group Size		0.091***	0.076***	0.080***
		(5.99)	(4.74)	(4.85)
Endogenous*Punishment			0.394***	0.216**
			(3.46)	(2.20)
Sanction-Free Groups				-0.414***
				(-8.07)
Constant	0.701***	-0.374	-0.299	-0.283
	(7.98)	(-1.52)	(-1.15)	(-1.06)
Observations	5121	5121	5121	5121
Adjusted R^2	0.005	0.361	0.393	0.435

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Dependent variable: *Efficiency*. t statistics in parentheses. Weights are inversely proportional to the variance of group contribution. 20 observations are excluded because of the zero-weight associated, due to no variance in the group contributions. Standard errors are clustered at group level

Table A4: Logit regression

	(1)	(2)	(3)	(4)	(5)
Endogenous	-0.426**	-0.191	0.237	-0.175	-0.149
	(-2.40)	(-1.09)	(1.14)	(-0.93)	(-0.63)
Group Size		-0.0808***			-0.107***
		(-4.03)			(-3.83)
% Cooperators			8.464***		9.784***
			(13.33)		(13.76)
Punishment				0.412	0.00766
				(0.83)	(0.01)
Endogenous*Punishment				-0.846	1.504***
				(-1.63)	(2.65)
Constant	1.877***	2.175***	-4.016***	1.829***	-4.511***
	(12.36)	(12.76)	(-8.48)	(11.37)	(-8.13)
Observations	15095	15095	15095	15095	15095
R^2					
Adjusted R^2					

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Dependent variable: *Matching*. t statistics in parentheses. Errors clustered at the individual level.

Table A5: Linear Probability Model - Free-riders

	(1)	(2)	(3)	(4)	(5)
Endogenous	0.120	0.120	0.104	0.028	-0.015
	(1.35)	(1.35)	(1.50)	(0.24)	(-0.28)
Group Size		-0.000			-0.001
		(-0.05)			(-0.23)
% Free-Riders			-1.218***		-1.392***
			(-7.28)		(-9.14)
Punishment				-0.239	-0.449***
				(-1.38)	(-3.63)
Endogenous*Punishment				0.529***	0.715***
				(2.74)	(4.98)
Constant	0.412***	0.412***	0.837***	0.437***	0.955***
	(5.91)	(5.88)	(10.38)	(4.82)	(13.74)
Observations	7901	7901	7901	7901	7901

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Dependent variable: *Matching* of free-riders with cooperative groups. Clustering at the paper, session, group and individual level. t statistics in parentheses

Table A6: Multilevel-Model

	(1)	(2)
Endogenous	0.082	-0.067
	(1.01)	(-0.97)
Punishment		0.106
		(0.73)
Endogenous*Punishment		0.496***
		(3.34)
Group Size		0.019***
		(12.26)
Constant	0.318***	0.218***
	(5.04)	(3.97)
Observations	7901	7901

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Dependent variable: Contributions of free-riders as % of initial endowment. Clustering at the paper, session, group and individual level. t statistics in parentheses

B Supplementary figures

Figure B1: First-round contribution as percentage of the initial endowment