universität innsbruck

Research Platform
Empirical and Experimental Economics

Ostracism and Theft in Heterogeneous Groups

Alexandra Baier, Loukas Balafoutas, Tarek Jaber-Lopez

Working Papers in Economics and Statistics

2021-19

University of Innsbruck
https://www.uibk.ac.at/eeecon/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
research platform “Empirical and Experimental Economics”
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 71022
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
https://www.uibk.ac.at/eeecon/wopec/

For a list of recent papers see the backpages of this paper.

Ostracism and Theft in Heterogeneous Groups

Alexandra Baier[a], Loukas Balafoutas[a] , and Tarek Jaber-Lopez[b,*]

[a] Department of Public Finance, University of Innsbruck.
[b] EconomiX, Université Paris Nanterre
* Corresponding author. 200, Avenue de la République 92000 Paris, France.
Email: tarek.jl@parisnanterre.fr. Tel:+ 33140977818

This version: 22.06.2021

Abstract

Ostracism, or exclusion by peers, has been practiced since ancient times as a severe form of punishment against transgressors of laws or social norms. The purpose of this paper is to offer a comprehensive analysis on how ostracism affects behavior and the functioning of a social group. We present data from a laboratory experiment, in which participants face a social dilemma on how to allocate limited resources between a productive activity and theft, and are given the opportunity to exclude members of their group by means of majority voting. Our main treatment features an environment with heterogeneity in productivity within groups, thus creating inequalities in economic opportunities and income. We find that exclusion is an effective form of punishment and decreases theft by excluded members once they are re-admitted into the group. However, it also leads to some retaliation by low-productivity members. A particularly worrisome aspect of exclusion is that punished group members are stigmatized and have a higher probability of facing exclusion again. We discuss implications of our findings for penal systems and their capacity to rehabilitate prisoners.

Keywords: ostracism, social dilemma, theft, rehabilitation, heterogeneous groups

JEL classification: C91, C92, K42

1. Introduction

Social exclusion or ostracism can be found in various aspects in life. Examples range from exclusion in sport clubs, bullying or cyber ostracism up to imprisonment as the harshest form of exclusion from society. [1] Group members can be excluded from any participation or interaction within a group when they do not comply with the rules or social norms. The practice of exclusion can be found in most societies in the world (Gruter & Masters, 1986) and its aim is in principle threefold: preventive, punitive, and corrective. Evidence from social psychology has documented diverse effects of social exclusion on individual behavior. On the one hand, ostracism can increase group conformity (Baumeister & Leary, 1995; Feinberg et al., 2014; Williams et al., 2000) , but on the other hand it also generates anger or lowers self-esteem (van Beest & Williams, 2006; Zadro et al., 2004). These negative emotions do not only occur with face-to-face ostracism, but also with ostracism on the internet: experiments on cyber ostracism show that participants report more negative feelings after being ostracized (Williams et al., 2000, 2002), and they even react sensitively to the slightest form of ostracism by a computer (Zadro et al., 2004).

In many economic experiments outlined in Section 2, exclusion can be used as a form of punishment and has been shown to help sustain cooperation and reduce free-riding in social dilemmas (Cinyabuguma et al., 2005; Maier-Rigaud et al., 2010; Sheremeta et al., 2009). However, relatively little research is available on the reintegration of excluded group members in economic experiments.[2] A number of important questions arise: are excluded individuals able to successfully reintegrate into their group after readmission? Do they display more pro-social behavior, or could exclusion backfire by leading to retaliation? Are group members willing to accept and reintegrate those returning from exclusion, or is exclusion associated with a persistent social stigma? Does the answer to the above depend on the group composition and heterogeneity? Motivated by these questions, we use a controlled laboratory experiment to investigate the relationship between exclusion and reintegration, both with respect to the behavior of excluded individuals and of remaining group members.

The experiment relies on a framework where subjects can either produce or steal from each other over a game horizon of 15 periods played in teams of four.[3] Hence, contrary to the large majority of economic experiments on the topic, we do not use a public goods game, but a different type of social dilemma that allows participants to engage in direct theft. Additionally, in each period we allow subjects to endogenously exclude team members based on a majority voting rule. In this environment, we create heterogeneity among the team members in terms of productivity. This leads to different incentives in the trade-off between stealing and producing, with low productivity members having stronger incentives to steal. The experimental design also includes a number of additional treatments, aimed at giving us a better understanding of behavior in this game. We add two treatments with homogeneous teams and one

[1] Ostracism defines the act of being ignored or excluded. This type of punishment goes back to ancient Athens and was used to establish a more secure and cohesive society by excluding individuals that might have threatened peace or democracy (Williams, 2007). Nowadays ostracism rather describes 'the practice of excluding disapproved individuals from interaction with a social group' (Hirshleifer & Rasmusen, 1989, p.89).

[2] One notable exception is Solda and Villeval (2020), which we discuss in some detail in Section 2.

[3] Throughout the paper and in order to avoid confusion, the word 'team' will refer to a group of four subjects who interact over the course of our experimental game, while the word 'group' will refer to one of two types within a team (low-productivity or high-productivity). The purpose of this terminological distinction is to avoid misunderstandings when we discuss terms such as 'in-group' or 'out-group' bias: such terms refer to bias driven by affiliation with a given type.

treatment without exclusion opportunities, in order to assess the role of introducing heterogeneity and exclusion into the environment. Moreover, to determine whether a pure in-group bias exists in stealing and voting decisions in this game (beyond the different monetary incentives faced by low and high productivity types), we also run two treatments with random assignment of members to minimal groups.

We contribute to existing literature in several ways. First, we address the above questions on exclusion and reintegration in a unified and rich setting, in which we have data on the stealing and voting behavior of excluded and non-excluded team members over time. Second, this study is – to our knowledge – the first to build heterogeneity into this setting, differentiating between advantaged (high-productivity) and disadvantaged (low-productivity) team members. The idea is to reproduce qualitative differences in socioeconomic status in the lab that are likely to affect exclusion (e.g., imprisonment) in reality. This feature of the design allows us to perform a detailed analysis on the effect of heterogeneity on stealing and exclusion, on in-group bias in stealing and voting and on how members of different types interact with each other in this setting. We use a number of control treatments in order to better understand the observed data patterns. Third, we work with an experimental game that in our view is very appropriate for capturing immoral behavior (theft) that can be punished by exclusion.

Our results show that exclusion can promote rehabilitation by discouraging antisocial behavior: members who re-enter their team after one period of exclusion steal significantly less than before. This key finding holds true for both types and in all treatments where exclusion is a feature of the design. However, we also find that there are at least two dark sides to exclusion. First, it leads to retaliation in the form of previously excluded members more frequently voting to exclude their peers when they re-enter the team and have voting rights. Second, controlling for a number of individual and team-specific factors, members who re-enter their team have a higher chance of being sent away again. This pattern suggests that there is a stigma associated with exclusion, hampering the successful reintegration of former transgressors. Our design also allows us to examine the interplay between high-productivity and low-productivity members within a team. We establish an in-group bias in stealing and in voting decisions, which is generally more prevalent among low productivity types and in heterogeneous – as opposed to minimal – groups. Finally, our control treatments reveal that introducing exclusion into the environment reduces theft, while heterogeneity within a team has no detectable effect.

As one of the most severe forms of social exclusion, imprisonment is used to enforce laws and social norms (Masters, 1984). However, little is known about the effect of imprisonment on reintegration and behavior. Statistical evidence from re-socialization and recidivism of former prisoners, for example, raises questions on the efficacy of imprisonment in deterring repeated offenses and promoting reintegration of prisoners after release. The Austrian Bureau of Statistics reports that, in 2016, 46.2% of all sentenced men in prison had committed a crime before.[4] Even higher numbers can be found for the United States, where 76.6% of inmates were re-arrested within 5 years after their release (Durose et al., 2014). These figures cast doubt on the preventive and corrective role of imprisonment and the extent to which it reduces crime, but they must be interpreted with caution since they suffer from problems of endogeneity and selection bias. At the same time, evidence on the relationship between imprisonment and reintegration is rather mixed.[5] Our work can contribute to this debate by presenting evidence from

[4] http://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/soziales/kriminalitaet/index.html)

[5] A number of studies using different methods point towards the direction of a positive effect of prison sentences in terms of crime reduction (see, for instance, Bhuller et al. 2016, 2018; Drago et al. 2009), while others suggest

a controlled experiment on how exclusion affects behavior after re-admission into a social group. The advantage of this approach is that we can isolate and cleanly identify the effects of exclusion in this context, while the obvious limitation relates to external validity and the fact that insights from the economic lab may not generalize to every aspect of the corrective system.

2. Experimental literature on ostracism

In general, ostracism is defined as 'being ignored or excluded' (Williams, 2002, 2007). In this work we will use the terms ostracism and (social) exclusion interchangeably to describe a broad class of situations and practices such as those mentioned in the introduction.[6] In laboratory experiments on ostracism, exclusion typically occurs either after interaction and separation from others within a group, or as a hypothetical consequence in the future (Williams, 2007). A popular experiment in psychology is a computer-based ball-tossing game where all players can be included or ostracized. This minimal ostracism paradigm was introduced by Williams (1997). There is evidence that social exclusion leads to prosocial and adaptive behavior but also to maladaptive (antisocial) reactions (see Bernstein & Claypool (2012) for an overview). Moreover, research from psychology shows that short time internet ostracism can have the same effect as short periods of face-to-face ostracism (Zadro et al., 2004). This finding supports the notion that the laboratory is an adequate environment for examining the effects of exclusion.

Existing literature in experimental economics shows that exclusion as a form of punishment can be a useful mechanism to solve social dilemmas and to foster cooperation (Akpalu & Martinsson, 2012; Charness & Yang, 2014; Cinyabuguma et al., 2005; Güth, Levati et al., 2007; Hirshleifer & Rasmusen, 1989; Lowen & Schmitt, 2013; Maier-Rigaud et al., 2010; Masclet, 2003; Neuhofer & Kittel, 2015; Solda & Villeval, 2020). These laboratory experiments differ in various aspects of the environment, like in the type of exclusion (irreversible or not), the length of exclusion and whether the length was exogenously or endogenously imposed. Cinyabuguma et al. (2005) show that contribution levels rise to almost 100% when excluded subjects have to play the remaining periods of a finite public goods game in another group with lower initial endowments. In Maier-Rigaud et al. (2010) punished subjects are completely and irreversibly excluded from all upcoming activities, and this mechanism also raises contribution levels to almost the full level. Similar to irreversible exclusion, exclusion for only one period increases contribution levels in the short run but cannot induce subjects to adhere to the norm of cooperation after elimination of the exclusion mechanism (Sheremeta et al., 2009). Neuhofer & Kittel (2015) make a direct comparison of irreversible versus one period exclusion, showing that a higher level of contribution is reached in the former case. In Davis and Johnson (2015) subjects are excluded from a social activity, i.e., a chat, rather than from the public good. Such 'social' ostracism is also effective in deterring free-riding.

In work closely related to ours, Solda & Villeval (2020) allow subjects to exclude each other through majority voting in a social dilemma and focus their analysis on the behavior of excluded subjects once they are re-admitted into their group. The length of exclusion can be exogenous and either short (one period) or long (three periods), or endogenously chosen by the group. Their results show that longer

that prison sentences promote criminal or antisocial behavior (Balafoutas et al., 2020; Bayer et al., 2009; Chen & Shapiro, 2007).

[6] Nevertheless, we note that social psychology partly differentiates between ostracism, social exclusion and rejection (Williams & Zadro, 2001).

exclusion increases cooperation more, but only when it is set exogenously. With an endogenous choice of exclusion length, a quicker reintegration limits retaliation. A recent study by Dannenberg et al. (2019) allows groups to choose by vote whether they want to implement an exclusion institution in a public goods game, much in the same spirit as Gürerk et al. (2006) who let participants vote for or against a punishment institution. Dannenberg et al. (2019) find that the share of groups who vote for the exclusion institution is substantial and increasing over time, and that the existence of this institution increases contributions by about 80%.

As subjects with heterogeneous characteristics might have different incentives to contribute in public good games (Buckley & Croson, 2006), there are a few studies investigating the enforcement of contributions in social dilemmas with heterogeneous populations. Reuben & Riedl (2013) show that punishment can overcome the free-riding problem in a public goods game with different types of heterogeneity, while Kingsley (2016) finds that the effectiveness of punishment in a public goods game with heterogeneous groups depends on the source of heterogeneity in endowments. Similarly, peer punishment has been shown to increase contributions when group members have heterogeneous productivity levels (Tan, 2008). Nevertheless, none of these studies consider ostracism or exclusion.

3. Experimental design, procedures and dataset

3.1. Framework

Our experimental setting is a modified version of the game used in Ahn et al. (2016, 2018). In all treatments, participants play in teams of 4 subjects and for 15 periods. They obtain an initial endowment of 10 tokens in each of the periods. In **Stage 1** in the respective period, this endowment must be fully allocated between the following two activities: production according to a given production function, or theft (see Figure A1 in the Appendix). In some of the treatments, subjects are randomly assigned the role of Type A or Type B (as explained in section 3.2.), which remains fixed until the end of the experiment. Tokens invested into production yield earnings according to one of the two production functions with decreasing marginal returns shown in Table 1. Theft in this environment means allocating tokens in order to steal ECU (experimental currency units) from another team member. Each token invested into stealing increases a subject's own earnings by 15 ECU and reduces the earnings of the team member targeted by theft by the same amount. This game represents a social dilemma: the social optimum is obtained when all subjects invest all of their tokens into production (given that theft is a purely redistributive activity and generates no social surplus), but individual incentives are such that it is more profitable to invest a certain number of tokens into theft. Hence, abstaining from theft can be seen as a contribution to the public good.

In **Stage 2**, subjects see an information table about the earnings of each team member in the current period (see Figure A2 in the Appendix). They also learn how much each team member earned through theft and how much through production, and how much each member lost to theft. However, they are not given any information about who stole from whom.[7] This procedure is repeated in all 15 periods and subjects have fixed player numbers that remain the same over all periods.

[7] This information can be deduced in the rare cases where only one subject allocated tokens to theft in a period.

Exclusion is based on a voting mechanism that is implemented in **Stage 2**, i.e., after allocation decisions have been made. With the information table described above available to them on the screen, all subjects are asked to indicate if they want to exclude one or more of the other players in the team. Following a majority voting rule, those team members who receive votes for exclusion from at least two other team members, are excluded for the upcoming period.[8] Excluded team members do not take allocation decisions, but they see a screen with the sentence "You have been excluded" for the duration of **Stage1**. Although they are not allowed to vote in **Stage 2**, they see the same information table as the active team members. After the one period of exclusion, they re-enter the team automatically. Non-excluded players cannot steal from excluded team members or vote to re-exclude them. A vote that results in a successful exclusion costs 1 ECU.

Table 1: Production functions of Type A and Type B

Token invested	Marginal units produced Type A	Marginal units produced Type B
1	24	18
2	22	16
3	20	14
4	18	12
5	16	10
6	14	8
7	12	6
8	10	4
9	8	2
10	6	0

Predictions for stealing in the stage game are straightforward and follow from the two production functions. Type A players have lower marginal returns from production than from stealing when investing more than 5 tokens into production. Thus, the optimal allocation for a payoff-maximizing player of this type is to invest 5 tokens into production and 5 into theft. For Type B players the marginal returns from stealing already exceed the marginal returns from production after the second token; therefore, the optimal allocation for players of this type is 2 tokens to production and the remaining 8 tokens to theft. When all players follow these predictions, expected earnings per period are 70 units for Type A and 64 units for Type B players.[9]

[8] The majority rule works as follows: if there are currently four members in a team, each member votes on whether or not he wishes to exclude each of the other three members. Hence, each member can receive up to three exclusion votes, and majority requires at least two votes. If there are currently three members in a team (because the fourth one is excluded for the period), then each member can receive up to two votes for exclusion, and hence majority coincides with unanimity and requires two votes. If there are only two members or one member in a team in a given period, no voting takes place.

[9] Each Type A member receives 100 units through production and 75 (=5x15) units through theft. At the same time, Type A is the target of 7 theft tokens on average (8 tokens from each of the Type B members plus 5 tokens from the other Type A member, divided by three possible targets). This leads to a period payoff of 175 – (7x15) =

Predictions for voting outcomes are less straightforward and depend on a number of assumptions regarding subjects' motivation and their expectations about future theft by each type. In particular, we assume that all agents are risk-neutral, selfish, and have no motivation other than maximizing their own payoffs. Hence, we do not consider in-group favoritism, retaliatory or reciprocal considerations, or other related behavioral motivations, thus assuming that subjects allocate their theft tokens randomly among the members that are currently in the group. We further assume that stealing decisions follow the predictions of 5 and 8 tokens invested into theft by Types A and B, respectively, and that all players correctly anticipate this.

From the perspective of an individual team member, excluding someone is costly for two reasons: first, due to the small cost of 1 ECU that is subtracted when the vote against another team member was successful, and second due to the fact that, with fewer players in the team, theft from other team members is more likely to be targeted at oneself in the following period.[10] On the other hand, excluding another team member can be individually beneficial since the excluded member cannot steal in the following period and the threat of theft to an individual is reduced. Based on these considerations and tradeoffs, we show in Appendix A.3 that Type B members are excluded in equilibrium, while Type A members are not. In particular, it is in the interest of both types to exclude Type B players, while a majority against Type A cannot be formed because it is not in the interest of low-productivity types to exclude the high-productivity ones. However, it is important to keep in mind that the above predictions rest on a number of restrictive assumptions and do not take into account a number of important factors. Behavior in our experiment is certain to depend on team dynamics and take into account observed outcomes over time (i.e., past voting and stealing decisions within the team). Hence, while we have included predictions on voting behavior as part of a complete description of the game, the experiment results should not be understood as a direct test of these predictions.

3.2. Treatments

HET: This is the main treatment employed in the experiment in order to examine exclusion and theft in heterogeneous teams. All subjects are randomly assigned to be either Type A or Type B, with two players of each type in each team. The two different production functions with decreasing marginal earnings are shown in Table 1. Type A players have a higher marginal productivity for every token: in this way, we introduce heterogeneity between team members along the productivity dimension, dividing teams into two advantaged, more productive members and two disadvantaged, less productive members. The type and productivity of all team members is common knowledge. In **Stage 1** of each period, subjects decide on the tokens they allocate to production and to stealing from other members. In **Stage 2**, team members can be excluded for one period via majority voting as described above. In the first interaction period, there is an **additional voting stage** prior to the Stage 1 (see Figure A3 in the Appendix). This first voting stage serves as setting in which exclusion decisions are not influenced by

70 units. Following the same reasoning, each Type B member receives 34 units through production and 120 (=8x15) units through theft, and loses on average 90 units to theft (6 tokens x 15 units per token). This leads to a per period payoff of 64. These calculations assume that all team members randomly divide their theft tokens among the other members in the team.

[10] An additional implication of exclusion is that it is not possible to steal from excluded members. However, this does not add to the costs of exclusion since one can always steal from remaining team members. One exception is the case of only one person left in the team. This possibility does not change our predictions, which state that only the two Type B members will be excluded and the two Type A members will remain in the team.

former stealing behavior, allowing us to document voting and preferences for exclusion in a cleaner context.

In addition to the main treatment HET that forms the backbone of our experiment and analysis, we conduct five treatments that serve as controls and as ways to achieve a more precise identification of various effects and mechanisms at work. These treatments are the following:

BASE: There is no voting in this treatment and **Stage 2** consists of the information table only. Hence, the difference between *BASE* and *HET* is that exclusion is possible in the latter, but not in the former. All other aspects (including heterogeneity among team members) are identical between the two treatments.

HOA: In this treatment, all subjects have the same production function, meaning that teams are homogeneous. All subjects are of Type A. Subjects still vote over exclusion in **Stage 2** (as in *HET*) but there is no heterogeneity in the team. We conduct this treatment as an additional control, to help us determine how heterogeneity in the team affects theft and voting behavior.

HOB: This treatment is the same as *HOA*, but all subjects have the production function for Type B.

MIA: We introduce minimal groups that do not differ in productivity. All subjects have the Type A production function and are randomly allocated into two color groups (yellow or red). This color allocation creates no economic difference for the two groups and only serves as a means of creating distinct identities within a team, in line with the minimal group paradigm (Chen & Li, 2009), which implies that the smallest group affiliation affects the in-group and out-group behavior. We conduct this treatment in order to investigate whether potential in-group favoritism in *HET* arises from distributional concerns due to the different production functions and resulting income levels between the two types, or due to group affiliation.

MIB: This treatment is the same as *MIA*, but all subjects have the production function of Type B.

3.3. Sample and procedures

The experiment was programmed in z-tree (Fischbacher, 2007) and conducted at the EconLab of the University of Innsbruck in November 2018 and March 2019. The subject pool consists of students from various academic backgrounds who were recruited using HROOT (Bock et al., 2014). In total 444 students (60.6% female) participated in 20 sessions and earned 17.42€ on average for approximately one hour of experiment. Before the experiment started, the instructions were read out loud and all subjects had to answer a number of control questions to ensure that they adequately understood the game rules and decisions.[11] At the end of the experiment and before receiving their payment, all participants completed a short questionnaire on socio-demographics.

The experiment consisted of 15 periods and the sum of payoffs from all periods was paid out privately and in cash directly after the experiment. We paid out all periods of the game in order to ensure that negative final payoffs were practically ruled out. A related consideration relates to negative wealth stands at the end of a period, and especially early on during the game: in principle, it is possible that a team member who (for any reason, including chance) receives a very large number of theft tokens has a negative payoff in a period, reducing his or her accumulated wealth stand compared to the previous

[11] Sample instructions for treatment *HET* are provided in Appendix A.1.

period. In order to preclude negative accumulated wealth stands in early periods, we endowed all team members with an initial endowment of 200 ECU at the beginning of the game.[12]

3.4. Descriptive statistics

Table 2 reports the mean levels of committed and received theft (in number of tokens invested into stealing), earnings from production and in total, exclusion ratio (the number of periods in which a subject gets excluded divided by the total number of periods), and votes cast and received (in number of votes between 0 and 3). The variables are disaggregated by treatment, by type, and – where appropriate – by previously excluded and non-excluded subjects. We will regularly refer to this table throughout the following sections.

Table 2: Summary statistics

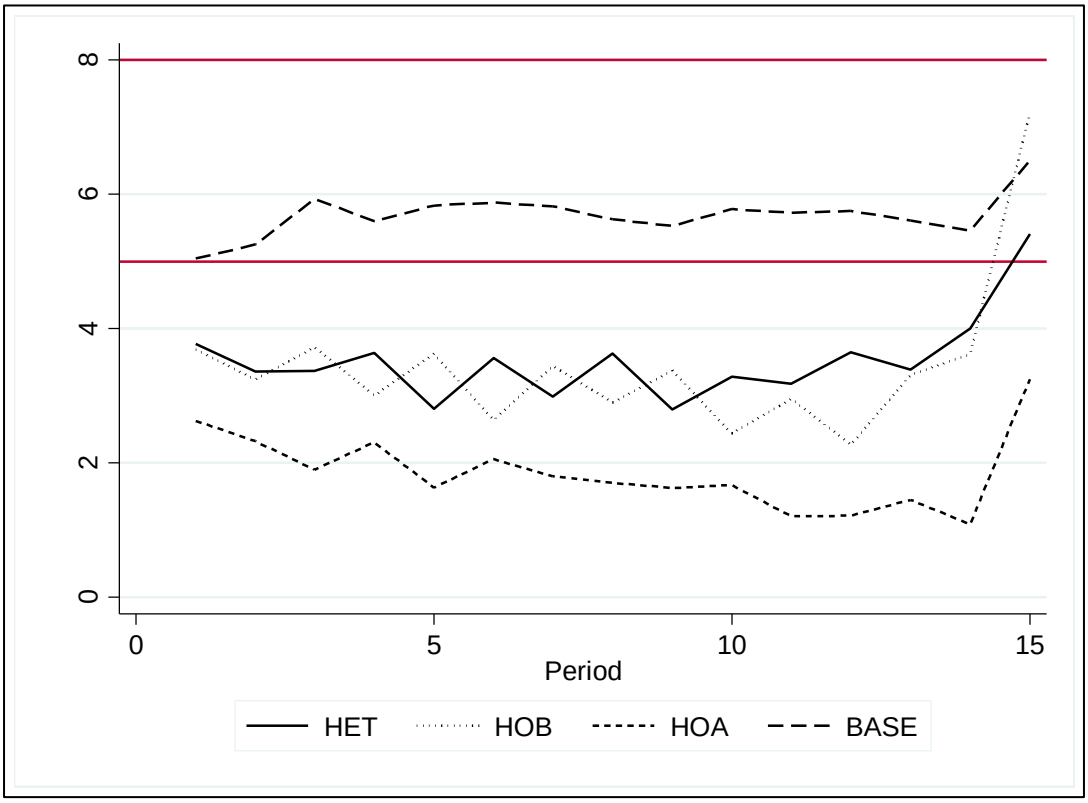
Treatment	**HET**		BASE		Homogeneous		Minimal	
Type	**A**	**B**	A	B	HOA	HOB	MIA	MIB
Tokens invested in Theft	**2.86**	**4.20**	4.76***	6.62***	1.86	3.48	1.66**	3.61
- Excluded Members	**3.51**	**5.21**	-	-	3.35	4.93	2.66*	4.64
- Non-Excluded Members	**2.54**	**3.63**	-	-	1.29*	2.88	1.37**	3.28
Received Theft	**3.97**	**3.05**	6.30***	5.08***	1.86***	3.48	1.66***	3.61
- Excluded Members	**4.34**	**3.41**	-	-	3.10***	4.74	2.58***	4.33
-Non-Excluded Members	**3.74**	**2.85**	-	-	1.37***	2.96***	1.39***	3.40***
Earnings from Production (in ECU)	**121.36**	**66.65**	99.94***	49.20***	131.49	71.97	134.74**	71.04
Total earnings (final in €)	**17.60**	**14.16**	16.90	16.06**	22.38***	13.15	23.92***	12.99
Exclusion ratio	**24.72%**	**28.47%**	-	-	22.25%	24.02%	18.52%	23.73%
Cast Votes	**1.14**	**1.12**	-	-	0.99	1.00	0.77**	1.06
- From Non-Excluded Members	**0.98**	**0.88**	-	-	0.71*	0.78	0.58***	0.84
- From Excluded Members after Readmission	**1.50**	**1.64**	-	-	1.81	1.54	1.53	1.61
Received Votes	**1.07**	**1.19**	-	-	0.99	1.00	0.77*	1.06
- Non-Excluded Members	**0.79**	**0.87**	-	-	0.71	0.74*	0.58**	0.82
- Excluded Members	1.77	1.79	-	-	1.80	1.64	1.51*	1.66
Number observations	720	720	540	540	1020	1020	1080	1020

Notes: Table reports mean values and stars indicate the results of Mann-Whitney's tests (henceforth MWU) comparing each type in each treatment to the corresponding type in treatment *HET* (in bold). *, **, *** indicates significance at the 10%, 5%, 1% level, respectively.. All results treat one average per team (over all periods) as one independent observation.

[12] In this respect, it is important to note that negative payoffs occurred in only 204 out of 6,660 subject-period observations, or 3.05% of the time, while the accumulated wealth never turned negative for any team member in any period.

Figure 1 shows the mean levels of tokens invested in theft over all 15 periods for all treatments. The prediction of 8 tokens invested into theft by Type B and 5 tokens by Type A are depicted via red lines. In line with these predictions, Type A players steal on average less than Type B players in all treatments (see the first row in Table 2; all comparisons between types are significant at the 5% level, MWU tests). Similar to Ahn et al., (2016, 2018) actual theft is substantially lower than predicted, although the difference is rather small in *BASE*, where the mean of tokens allocated to theft (5.69) is relatively close to the mean prediction of 6.5 tokens for both types pooled.

Figure 1: Evolution of average tokens invested in theft over time by treatment



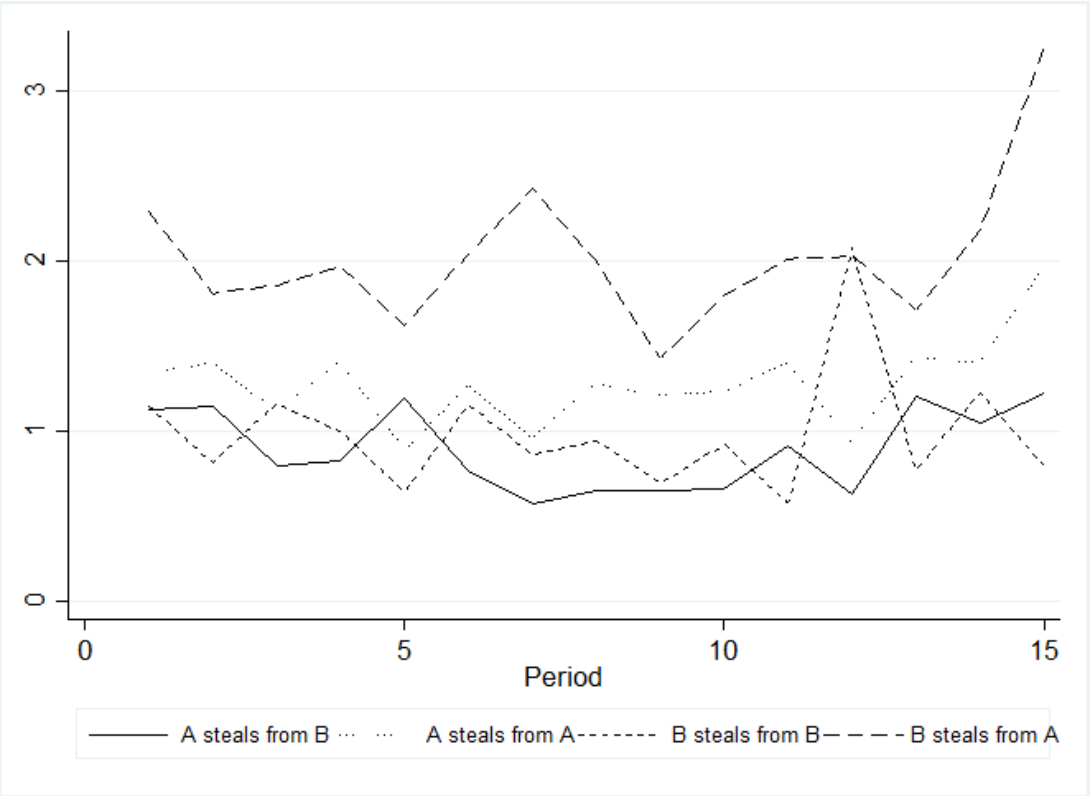
4. Behavior in heterogeneous teams with exclusion

4.1. Stealing behavior

Subjects in treatment *HET* invest 3.52 tokens into theft on average. In line with the incentives arising from the different production functions, Type A subjects steal significantly less than Type B subjects (2.86 vs. 4.20; p=0.003, Wilcoxon signed-rank test, henceforth WSR).[13] Figure 2 shows the mean number of tokens invested in theft, by type and targeted type. Type A players invest significantly more tokens in order to steal from Type B than from Type A players (1.31 vs. 0.91 tokens, p=0.002, WSR). The size of this bias is slightly stronger for players of Type B, who invest on average 2.05 tokens to steal from Type A but only 1.03 tokens to steal from other Type B players (p<0.001, WSR).

[13] Throughout the analysis, all non-parametric tests are based on conservative definitions of what constitutes an independent observation. In between-subjects comparisons, we treat one average per team (over all periods) as one independent observation. In paired comparisons entailing related samples, we treat the average behavior of all subjects in a sample and over all periods within a team as one independent observation (e.g., comparing Type A vs. Type B members, or comparing members who have just been re-admitted into the group vs. the rest).

Figure 2: Stealing by Type in *HET*



Our central question of interest with respect to stealing behavior in this setting is whether exclusion can discourage theft. As indicated in the second and third row of Table 2, when comparing subjects who spent the previous period in exclusion and re-enter the team against those who were not excluded in the previous period in *HET*, we find that the former choose higher levels of theft on average. This is true for Type A (3.51 vs. 2.54 tokens, $p<0.001$, MWU) and for Type B players (5.21 vs. 3.63 tokens, $p<0.001$, MWU). Clearly, however, these findings do not imply a causal impact of exclusion on theft, since excluded subjects may be more prone to theft in the first place, meaning that exclusion cannot be treated as exogenous with respect to theft. We control for this endogeneity using regression techniques in order to estimate the true effect of exclusion on theft and examine in more detail the various factors that drive stealing decisions in heterogeneous teams.

Table 3 presents multilevel regression models for *HET* with random effects on the subject and team level, with the number of tokens invested into stealing as the dependent variable.[14] The explanatory variables are defined as follows. The dummy *TypeA* takes value 1 if the subject was of Type A and 0 otherwise. $Exclusion_{t-2}$ is also a dummy variable, indicating whether the subject was excluded two periods before and re-admitted to the team in the current period or not. *MeanTheft* is the average of tokens allocated to theft per active period. This variable thus measures a subject's average stealing behavior, up until the previous period *(t-1)*. *MeanTheft_Received* is the average of theft suffered by a subject over all periods when the subject was active in the team (i.e., not excluded), up until *t-1*. For both averages, as well as for $Exclusion_{t-2}$ we add the interactions with Type A subjects, allowing for

[14] We use a multilevel model to allow for different sources of variation in the data. Thus, we account for the effect of the specific team as well as for the individual subject. A Tobit regression (dependent variable left-censored at 0 and standard errors clustered at the team level) can replicate the main results, most importantly the significant negative effect of $Exclusion_{t-2}$. See Table A1 in the Appendix.

heterogeneous effects by type. Further controls include *Period*, *Team size*$_t$ defined as the number of non-excluded subjects in the team in the respective period, and a *Female* dummy.

Table 3: Determinants of stealing in *HET*

	(1)	(2)	(3)	(4)
Dependent variable	Theft	Theft	Theft$_t$ - Theft$_{t-2}$	Theft$_t$ - Theft$_{t-2}$
TypeA	-0.255*	-0.174	-0.192	-0.468
	(0.145)	(0.294)	(0.182)	(0.370)
Exclusion$_{t-2}$	-0.386***	-0.356**	-1.126***	-1.367***
	(0.131)	(0.179)	(0.186)	(0.260)
Exclusion$_{t-2}$ *x TypeA*		-0.053		0.476
		(0.248)		(0.363)
MeanTheft	1.012***	1.031***	0.009	0.039
	(0.035)	(0.045)	(0.045)	(0.059)
MeanTheft x TypeA		-0.058		-0.056
		(0.071)		(0.093)
MeanTheft_Received	0.097**	0.080	0.032	-0.023
	(0.040)	(0.063)	(0.051)	(0.087)
MeanTheft_Received x TypeA		0.036		0.087
		(0.079)		(0.108)
Period	0.068***	0.067***	0.082***	0.082***
	(0.013)	(0.013)	(0.022)	(0.022)
Team size$_t$	0.437***	0.440***	0.389***	0.389***
	(0.072)	(0.072)	(0.118)	(0.119)
Female	0.130	0.141	-0.112	-0.106
	(0.147)	(0.146)	(0.173)	(0.174)
Constant	2.130***	2.180***	1.530***	1.397***
	(0.318)	(0.348)	(0.499)	(0.532)
Observations	999	999	802	802
Number of teams	24	24	24	24

Notes. Multilevel regressions, with subject and team random effects. Standard errors are clustered at team level. Dependent variable in (1) and (2): Theft tokens invested by subject *i* in period *t*. Dependent variable in (3) and (4): Theft tokens invested by subject *i* in period *t*, minus tokens invested by subject *i* in period *t-2*. Independent variables described in text. *, **, *** indicates significance at the 10%, 5%, 1% level, respectively.

The first specification in Column 1 shows a regression on the explanatory variables *TypeA*, *Exclusion*$_{t-2}$, *MeanTheft*, *MeanTheft_Received* and the controls *Period*, *Team size*$_t$ and *Female*. The negative coefficient of *TypeA* confirms our finding from the descriptive analysis that, on average, Type A players steal less than Type B players. Our key variable of interest, *Exclusion*$_{t-2}$, has a negative and highly significant coefficient. This goes against the descriptive analysis presented above, the reason for this reversal being that in the regressions we can control for past stealing behavior. Hence, we find that the experience of exclusion significantly decreases individual theft. This indicates a disciplining effect of exclusion on subjects who are re-admitted into their team and delivers an optimistic message regarding the possibility of successful reintegration. *MeanTheft* and *MeanTheft_Received* are both positive and significant. Thus, unsurprisingly, there is positive correlation over time in individual theft decisions, while subjects who have experienced a lot of theft before steal significantly more. This indicates the presence of retaliatory behavior. The control variables indicate that there is a positive time trend in stealing and that theft increases with more active players in the team, while we find no gender differences in stealing.

In Column 2 we additionally interact $Exclusion_{t-2}$, *MeanTheft* and *MeanTheft_Received* with the *Type A* dummy to investigate whether the aforementioned effects vary by type. We find that the disciplining effect of exclusion does not differ by type (given the insignificant interaction term $Exclusion_{t-2}$). This effect is captured by $Exclusion_{t-2}$ for Type B, and by the joint coefficient ($Exclusion_{t-2}$+$Exclusion_{t-2}$ *x TypeA*) for Type A (p=0.025, F-Test). The significant positive effect of *MeanTheft* also does not vary by type. *MeanTheft_Received* has no significant effect on the stealing decision for Type B, but it does for Type A (p=0.021, F-test). The overall effect of being Type A loses its significance (p=0.402, F-test).[15]

There is still a possible endogeneity issue regarding the stealing decision: excluded members are on average more prone to theft in the first place, hence the coefficient of the variable $Exclusion_{t-2}$ may be biased. In this respect, we first note that such a positive relationship would lead to an upwards bias of the coefficient, implying that the results shown in columns (1) and (2) may be underestimating the true rehabilitation effect of exclusion. To investigate this possibility, we follow Solda and Villeval (2020) and present in columns (3) and (4) of Table 3 additional regressions where the dependent variable is the evolution of the stealing between a given period and two periods before. The coefficients of $Exclusion_{t-2}$ in columns (3) and (4) are again highly significant, and they are almost three times higher than in the first two columns where we do not correct for possible endogeneity. Hence, the evidence on the positive rehabilitation effect of exclusion appears conclusive. *Type A* is no longer significant in these specifications. The variable controlling for past stealing behavior (*MeanTheft*) also loses significance, which is hardly surprising since we now control for the propensity to steal by using an appropriate dependent variable. The interaction terms with *TypeA* in Column 4 show that there is no significant difference in the effects of our main explanatory variables between Type A and Type B players. Finally, the positive time trend and the number of subjects in the team retain their significance.

As a robustness check and in order to see whether the effect of exclusion depends on how well a given team is functioning, we also run regressions where we split the sample by the median theft level. Median theft in *HET* lies at 3.5 tokens invested into stealing per team and per period. All teams with a below median theft level are thus classified as low theft teams, all others as high theft teams. In the Appendix (Table A2) we present the regressions on the evolution of theft (analogous to Table 3, column 3) separately for high and low theft teams. We find that being excluded significantly decreases stealing level for both types, in low as well as in high theft teams. Taking all above findings above into account, we formulate our main result regarding the effect of exclusion on stealing behavior.

Result 1: *Exclusion has a disciplining effect on stealing behavior in heterogeneous teams: subjects who are re-admitted into their team steal significantly less than they did before they were excluded, regardless of their type.*

Besides the effect of exclusion on the stealing decision from the perspective of the perpetrator (i.e., on the number of tokens allocated by a given subject to theft), we are interested in how theft suffered from the perspective of the 'victim', measured as the number of theft tokens that a given subject receives, is affected by exclusion. This relates to our research question on how team members treat those who return from exclusion. Table 2 suggests that subjects who have been excluded and return to their team experience more theft than those subjects who were not excluded in the period before (4.34 vs.

[15] As a robustness check, we have estimated the Table 3 regressions excluding the last period, since theft increases sharply in the last period (see Figure 1). All results hold in direction and magnitude.

3.74 for Type A and 3.41 vs.2.85 for Type B; both differences are not statistically significant according to MUW tests). In the Table 4 regressions, the dependent variable is thus received theft, while the explanatory variables are the same as in Table 3. The regression results reveal that, controlling for theft committed and suffered by a given subject in the past, exclusion does not affect the level of current received theft. This holds equally true for both types. It thus appears that, when we consider the stealing dimension, previously excluded members who re-enter the team are not subject to additional punishment and retaliatory behavior by the rest of the team.

Result 2: *We find no retaliation in terms of increased theft against previously excluded members who re-enter the team.*

Table 4: Determinants of received theft in *HET*

VARIABLES	(1) Theft received	(2) Theft received
TypeA	-0.201	-0.028
	(0.193)	(0.396)
$Exclusion_{t-2}$	-0.092	-0.123
	(0.172)	(0.235)
$Exclusion_{t-2}$ *x TypeA*		0.067
		(0.326)
MeanTheft	0.040	0.080
	(0.046)	(0.060)
MeanTheft x TypeA	1.201***	1.185***
	(0.053)	(0.084)
MeanTheft_Received		-0.095
		(0.095)
MeanTheft_Received x TypeA		0.035
		(0.105)
Period	0.075***	0.075***
	(0.017)	(0.017)
$Team\ size_t$	0.892***	0.893***
	(0.094)	(0.094)
Female	-0.095	-0.081
	(0.196)	(0.197)
Constant	-4.281***	-4.397***
	(0.418)	(0.459)
Observations	999	999
Number of teams	24	24

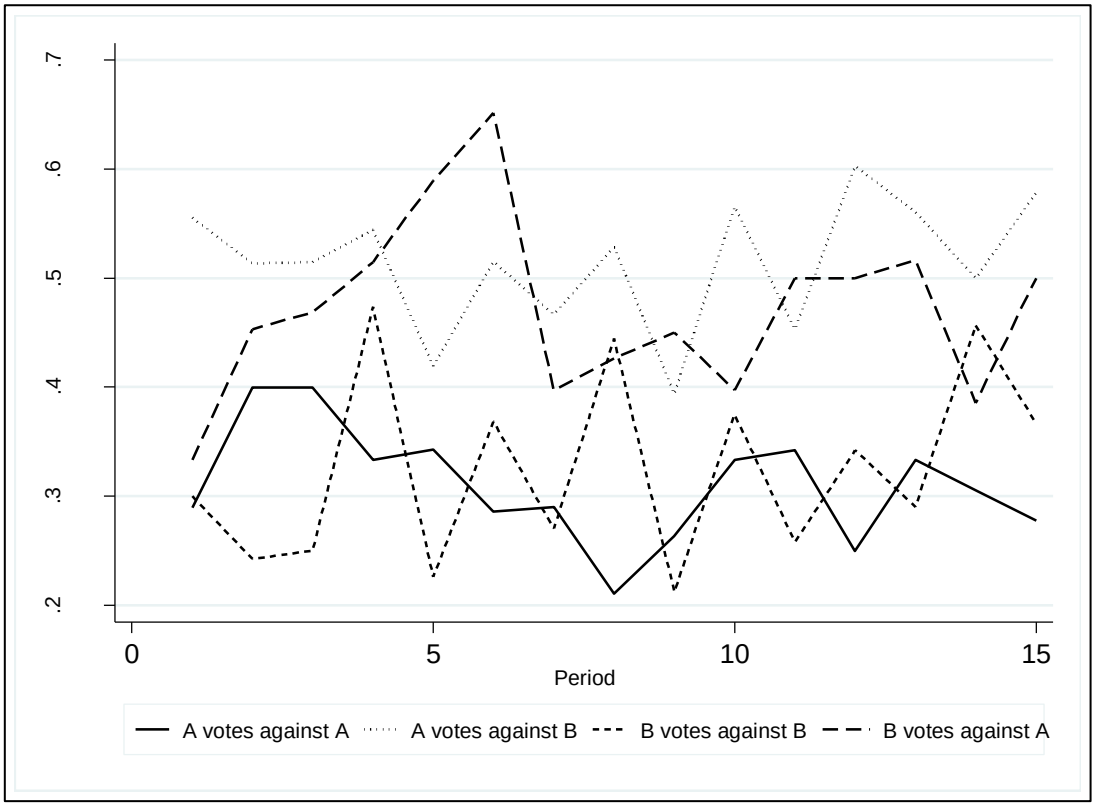
Notes. Multilevel regressions, with subject and team random effects for HET treatment. Standard errors are clustered at team level. Dependent variable: Received theft tokens by subject *i* in period *t*. Independent variables described in text. *, **, *** indicates significance at the 10%, 5%, 1% level, respectively.

4.2. Voting behavior

We now turn to behavior and outcomes in the second stage of each period in treatment *HET*. In this stage, subjects vote on whether they want to exclude each of the other team members for one period. Accordingly, there are two possible variables to consider: the occurrence of exclusions measured by means of a binary exclusion variable, and the number of votes cast. In the remainder of this section we place our focus on votes cast, but for completeness we also report here the mean exclusion rate, which is 26.6%. Type B players are excluded slightly more frequently than Type A players (28.5% vs. 24.7%), but this difference is not statistically significant (p=0.235, WSR).

Figure 3 shows the average of votes cast in favor of exclusion split by type and targeted type in *HET*, defined as the share of potentially possible votes. Overall, the average of cast votes per active team member does not differ significantly by type (0.45 for Type A vs. 0.46 for Type B; p=0.753, WSR). However, taking into account the type of the targeted player, we find strong evidence of in-group favoritism: Type A players vote more against Type B players (0.33vs. 0.53; p=0.006, WSR), while Type B players vote more against Type A players (0.35 vs. 0.49; p=0.018, WSR).

Figure 3: Mean of votes by type and targeted type over periods in *HET*



Voting decisions are affected not only by a potential in-group bias, but also – and arguably more strongly so – by the observed behavior of other team members over time. Therefore, a cleaner test for the presence of an in-group bias can be made using the voting decisions in the initial stage *before* the first period of the game, and hence before subjects have taken any allocation decisions. This initial voting stage was implemented precisely for the purpose of allowing us to assess in-group bias in voting. Figure A5 in the Appendix displays the average of votes cast by subject type and by targeted type in this

first voting stage. We find no in-group bias in the decisions of Type A players (on average 0.31 votes against Type A vs. 0.33 votes against Type B p=0.977, WSR), while for Type B players there appears to be some bias against the other type; this, however, is not significant (0.27 votes against Type A vs. 0.17 votes against Type B, p=0.103, WSR).

Result 3: *Taking the entire interaction into account, both types display in-group favoritism in their voting decisions. However, there is no significant group bias in voting in the initial voting stage.*

One final observation from Table 2 is that subjects in *HET* tend to cast more votes in the period when they are re-admitted into the team after exclusion, compared to the rest of the sample (1.57 vs. 0.93 votes; p=0.004, WSR). This is true both for Type A and for Type B players (p<0.018 for both comparisons). This hints towards retaliation in the voting dimension, as discussed also in Solda & Villeval (2020). Nevertheless, an in-depth examination of the various motives behind voting decisions requires taking into account several forces and dynamics over the course of interaction. For this purpose, we present in Table 5 a series of multilevel regressions on the votes that a player gives (in the first two columns) as well as on the votes a player receives (in the third and fourth column) in treatment *HET*. The dependent variable is the share of votes received (or given), defined as the number of votes a player receives (or gives) divided by the maximum number of votes possible in a given period in order to account for the fact that the number of active members varies over time in a team. Again, the hierarchical model includes random effects at the group and at the subject level in all specifications. The explanatory variables are largely the same as in the theft regressions, albeit with some notable differences: we now use committed and received theft in the current period and not the average over all previous active periods, given that theft in the current period is expected to be the main driver of voting decisions and this information is displayed on the screen where voting takes place. In addition, the number of members active in the team is now omitted (since it is part of the dependent variable).

The most important observation from the first two columns is the positive and significant coefficient for $Exclusion_{t-2}$, which means that excluded subjects who return to the team give significantly more votes than non-excluded subjects. This confirms the findings in the descriptive analysis. However, an important observation is that this effect is driven by Type B members only: the significant interaction term $Exclusion_{t-2}$ *x TypeA* in the second column shows that Type A players react differently to exclusion, and their reaction is close to zero and not significant (joint coefficient $Exclusion_{t-2}$ + $Exclusion_{t-2}$ *x TypeA* = 0.018, p=0.592, F-test).

In terms of further predictors of votes cast, we find that Type A players do not cast significantly more or fewer votes than Type B players. Higher theft in the current period slightly increases the number of votes, which means that those individuals who steal more are on average also more likely to vote to exclude others – a pattern reminiscent of the well-documented phenomenon of antisocial punishment in social dilemmas (see, e.g., Cinyabuguma et al., 2006; Hermann et al., 2008; Nikiforakis, 2008). Being a victim of much theft also increases a subject's votes to exclude others, in line with a retaliation story but also with an effort of victims to protect themselves from theft in the next period by excluding thieves from the team. The significant interaction terms in the second column show that the two types react differently to committed and suffered theft. First, Type A players' votes are not significantly influenced by their committed theft (joint coefficient *Theft + Theft x TypeA* =-0.007, p=0.908, F-test). Second, the increase in voting after suffering a lot of theft is driven by both types, but it is significantly stronger for

Type A. As a robustness check, in the Appendix (Table A4) we also present regressions with the average of committed and suffered theft as right-hand side variables instead of the theft levels in the current period. The main results (notably the effect of exclusion on voting) still hold. Hence, while Result 2 reports no retaliation of excluded subjects in the stealing dimension, we do find evidence of retaliatory voting.

Result 4: *Low productivity (Type B) players retaliate by voting more after returning from exclusion. In addition, both types cast more votes after experiencing a lot of theft.*

Table 5: Determinants of voting in *HET*

VARIABLES	(1) Votes Cast	(2) Votes Cast	(3) Votes Received	(4) Votes Received
TypeA	-0.006	0.024	0.053***	0.057
	(0.041)	(0.060)	(0.020)	(0.037)
$Exclusion_{t-2}$	0.079***	0.136***	-0.035	-0.044
	(0.024)	(0.033)	(0.022)	(0.030)
$Exclusion_{t-2}$ *x TypeA*		-0.118**		0.017
		(0.047)		(0.041)
Theft	0.009*	0.016**	0.043***	0.044***
	(0.005)	(0.006)	(0.004)	(0.005)
Theft x TypeA		-0.017*		-0.001
		(0.010)		(0.007)
Theft_Received	0.025***	0.013**	0.002	0.000
	(0.004)	(0.006)	(0.003)	(0.005)
Theft_Received x TypeA		0.021***		0.003
		(0.008)		(0.006)
Cumulative_Exclusions			0.122***	0.127***
			(0.008)	(0.010)
Cumulative_Exclusions x TypeA				-0.010
				(0.012)
Period	-0.003	-0.003	0.030***	0.030***
	(0.002)	(0.002)	(0.003)	(0.003)
Female	-0.088*	-0.085*	0.032	0.033
	(0.046)	(0.046)	(0.022)	(0.022)
Constant	0.384***	0.370***	0.266***	0.263***
	(0.053)	(0.057)	(0.035)	(0.038)
Observations	964	964	964	964
Number of teams	24	24	24	24

Notes. Multilevel regressions, with subject and team random effects. Standard errors are clustered at team level. Dependent variable in (1) and (2): Votes cast by subject *i* in period *t* defined as the share of all potentially possible votes according to the number of active player in the team. Dependent variable in (3) and (4): Share of potentially possible votes received by subject *i* in period *t*. Independent variables described in text. *, **, *** indicates significance at the 10%, 5%, 1% level, respectively.

The third and fourth column of Table 5 examine the determinants of voting from the perspective of votes received, in order to illustrate the characteristics of those team members who are more likely to be excluded. In addition to theft, these specifications also include the cumulative sum of exclusions until the current period, *Cumulative_Exclusions*. This variable indicates how often a given subject has been excluded before. The key thing to note here is that the coefficient of $Exclusion_{t-2}$ is insignificant, while the coefficient of *Cumulative_Exclusions* is positive and highly significant. This means that subjects who have been excluded more often in the past receive more votes in the current period (and hence are more likely to be excluded again), controlling for their current theft. We interpret this result as a stigma associated with past exclusion.[16] This is quite worrying, since it implies that the team itself poses obstacles to a successful reintegration of formerly excluded members, potentially creating a vicious circle of exclusions and retaliation. This effect does not significantly vary by type, as seen by the insignificant interaction of $Exclusion_{t-2}$ with *TypeA* in column (4).[17] We further find that subjects who stole more in the current period receive significantly more votes regardless of their type and also that, contrary to the static prediction from 3.3., players of Type A are more likely to receive exclusion votes, ceteris paribus.

Result 5: *Subjects of both types with a higher cumulative number of past exclusions are more likely to be excluded in the present period, ceteris paribus.*

5. Additional treatments: Baseline, homogeneous teams and minimal groups

5.1. Treatment *BASE*: The role of introducing exclusion

We have already presented a thorough econometric analysis on the question of how experienced exclusion impacts theft, using data from heterogeneous teams. In addition to this, the baseline treatment without exclusion opportunities (*BASE*) allows us to assess how the possibility of exclusion affects theft, i.e., how overall theft levels change when moving from an environment where exclusion is impossible to one where it is part of the game. We find that, on average, subjects invest significantly more tokens into theft in treatment *BASE* than in treatment *HET* (3.53 tokens in *HET* vs. 5.69 in *BASE*; $p<0.001$, MWU). As shown in Table 2, this result holds for both types of subjects.[18] This supports findings from other studies that the threat of punishment reduces anti-social behavior.

We also briefly report how the two treatments compare in terms of efficiency. Efficiency in our experiment is measured in terms of earnings of all team members. In this respect, Table 2 shows that the average earnings for Type A players increase slightly when exclusion is possible, from 16.90€ to 17.60€, although the difference between *BASE* and *HET* is not significant. On the contrary, the payoffs for Type B decrease significantly from 16.06€ in *BASE* to 14.16€ in *HET*. Thus, the possibility of exclusion harms only Type B players. Overall, we confirm the finding from Solda and Villeval (2020)

[16] This is made possible by the fact that subjects can be identified by means of their fixed player numbers within a team.

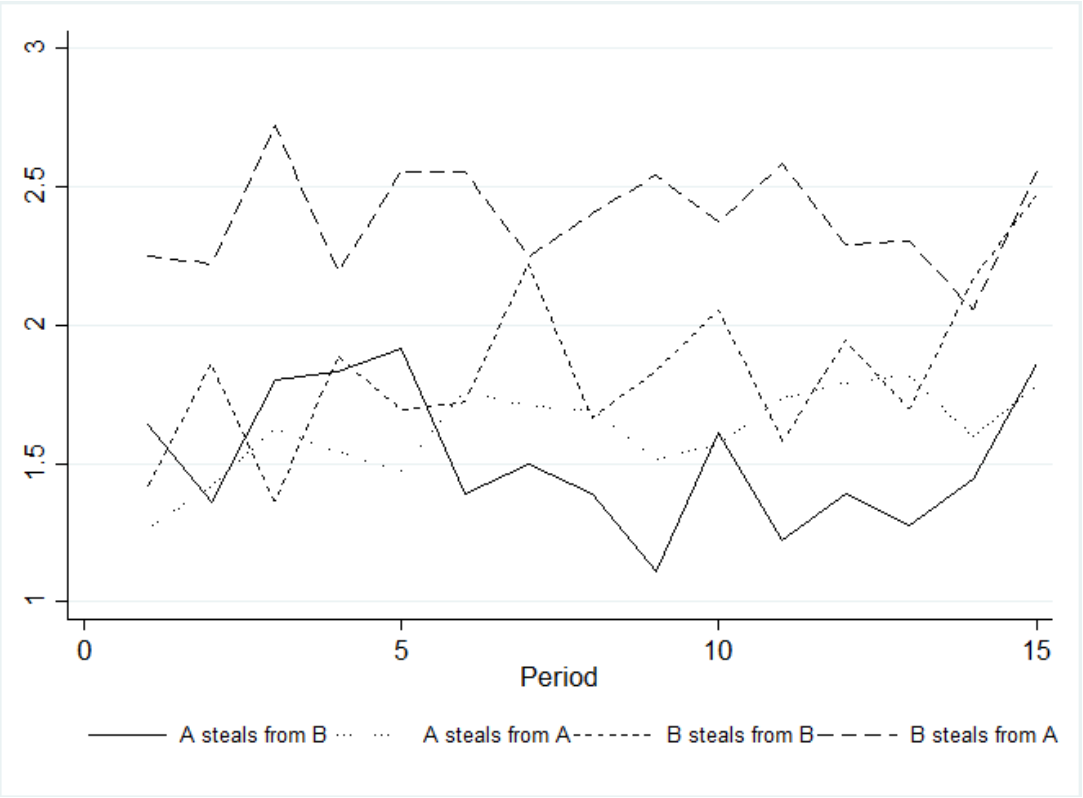
[17] This finding is robust to replacing the current theft variables with average committed and received theft. The regressions are shown in the Appendix (columns 3 and 4 in Table A4).

[18] The magnitude of this effect is slightly higher for Type A (1.90 tokens) than for Type B members (2.42 tokens). However, a linear regression (available upon request) reveals that the difference-in-differences is not significant.

that exclusion does not systematically harm efficiency, since total earnings are very comparable between the two treatments (15.88 in *HET* vs. 16.48 in *BASE*, p=0.357, MWU).

The next issue where the baseline treatment is informative relates to group favoritism across types. For this purpose, we disaggregate theft by targeted type. Figure 4 shows stealing decisions for treatment *BASE* over all periods, split by type and by targeted type. We see that Type A players steal slightly (but not significantly) more from Type B than from other Type A players (1.62 vs. 1.52 tokens; p=0.965, WSR). For Type B players we find a significant difference between the targeted groups, as they steal more from Type A than from other Type B players (2.39 vs. 1.84 tokens invested; p=0.012, WSR).

Figure 4: Stealing by Type in *BASE*



This documented gap over the entire course of the game in theft may be driven by a pure in-group bias, by experiences and dynamics within a team over time, or by a combination of the two. Considering decisions only in the first period of the game in *BASE* allows us to establish whether there is in-group favoritism in stealing, or if team dynamics that evolve over time account for the differences described above. Figure A4 in the Appendix shows theft from both types and by targeted type in the first period in *BASE*. For Type A, we find that theft against other Type A players is slightly *higher* than theft against Type B players on average, although the difference is not significant (1.64 vs. 1.26, p=0.146, WSR). For Type B players we find clear evidence of discrimination against Type A (2.25 vs. 1.42, p=0.020, WSR).

One final question related to in-group bias is whether it arises from group identity per se, or whether it is specific to the distinction between an advantaged (high-productivity) and a disadvantaged (low-productivity) type. Our two treatments with minimal groups (*MIA* and *MIB*) allow us to address this question. We do so in section 5.3

5.2. Treatments HOA and HOB: Behavior in homogeneous teams

One interesting question is whether and how heterogeneity per se affects stealing decisions. To this end, we compare treatment *HET* against the two treatments with exclusion but without heterogeneity (*HOA* and *HOB)*, noting that the monetary incentives for each subject type to invest tokens into theft are identical in the heterogeneous and in the respective homogeneous treatment (5 tokens for Type A and 8 tokens for Type B). Type A players steal more in *HET* (2.86 tokens) than in the homogeneous treatment *HOA* (1.86 tokens). Similarly, Type B players steal more in *HET* (4.20 tokens) than in *HOB* (3.48 tokens). However, in both cases, the treatment differences are not statistically significant (p=0.107 for Type A and p=0.137 for Type B, MWU). This suggests that team heterogeneity – along the productivity dimension, as implemented in our experiment – does not increase theft in the team.[19]

We can also use the data from the two homogeneous treatments to assess whether the disciplining effect of exclusion on theft, documented in Result 1, carries over to settings with homogeneous teams. In the Appendix (Table A3) we present versions of the Table 3 regressions in teams with the possibility to exclude others and members that are either all of Type A (column 1) or Type B (column 2). In line with the findings presented in section 4.1, the coefficient on $Exclusion_{t-2}$ is negative, very sizeable and highly significant for both specifications.

In terms of exclusion, we report that mean exclusion rates for both types pooled are somewhat higher in *HET* (26.6%) than in the homogeneous treatments *HOA* (22.3%) and *HOB* (24.0%). However, none of these comparisons are statistically significant ($p>0.39$, χ^2 tests). Hence, from an aggregate perspective, heterogeneous teams are not associated with more (or less) exclusion compared to homogeneous ones. Another thing that does not vary between the heterogeneous and homogeneous setting is the retaliatory behavior by recently excluded subjects that we documented in section 4.2. (Result 4). This kind of behavior is visible in both homogeneous treatments, with re-admitted subjects casting more votes than non-excluded subjects do on average (1.81 vs. 0.71 votes in *HOA* and 1.54 vs. 0.78 in *HOB*; $p<0.002$ for both comparisons, WSR). Hence, retaliation does not hinge on heterogeneity in the team, but rather on the actual event of experiencing exclusion.

5.3. Treatments MIA and MIB: Bias in minimal groups

As already explained in section 3.2., the motivation for running the minimal group treatments has been to gain a better understanding of what drives potential in-group bias. Results 2 and 3 show that there is in-group bias in stealing – and to a lesser extent voting – decisions. This bias could be driven by distributional preferences and the fact that the two types have different productivities, such that Types A are advantaged and have substantially higher earnings on average (see Table 2). The perception of unfairness can be particularly strong given the random assignment of subjects to types. Alternatively, group bias could be the product of a pure preference for members of the in-group, regardless of income differences. The minimal group treatments can help us to say something on the relative weight of the

[19] In regards to these comparisons, it must be kept in mind that the behavior of a particular type may differ between the heterogeneous and homogeneous treatments due to several reasons, related to team dynamics and experiences over the course of the game as well as to the different average productivity and income levels (which are highest in *HOA*, lowest in *HOB* and intermediate in *HET*).

two mechanisms and establish a true in-group bias, since productivities here are held constant and either high (in *MIA*) or low (in *MIB*), while different group identities are still established. In these treatments, differences in theft across targeted group will hint to a true bias and cannot have a distributional or fairness background.

Comparing the average levels of in-group and out-group theft in the two minimal group treatments, we find that in both there is a difference between stealing from the in-group and from the out-group. In *MIA* the average of tokens invested into theft is 0.59 tokens targeted at the in-group and 0.75 tokens targeted at the out-group. This difference, however, is not significant (p=0.139, WSR). A significant group bias can be found in the Type B minimal groups (treatment *MIB*), with significantly less stealing from the in-group than from the out-group (1.21 tokens vs. 1.6 tokens, p=0.050, WSR). Comparing this to the bias of Type B players in *HET* (who invest 2.05 tokens into theft from Type A and 1.03 tokens into theft from Type B), we see that the size of this bias is much smaller in *MIB* than in *HET* in absolute terms (0.39 versus 1.02 tokens) and in relative terms (increase of 32% versus 99%). This result indicates that fairness concerns appear to be a major force behind the stealing decision for type B players. The group affiliation introduced via different colors is enough to generate salient group identities and a bias against the out-group, but the size of this bias is about one third compared to the treatment where heterogeneity is implemented along the productivity dimension.

Turning to voting decisions, subjects in *MIA* cast 0.28 votes against in-group members and 0.32 votes against out-group members on average. This difference is not significant (p=0.117, WSR). In *MIB*, average votes against in-group members are significantly lower than against out-group members (0.37 vs. 0.41 votes, p=0.047, WSR). Hence, we establish that subjects in *MIB* vote more to exclude team members of the different color, even though there is no difference between the two in terms of productivity or income. We note, however, that the size of this bias in MIB (0.04 votes, or an increase of 11%) is much smaller than the bias of Type B players in HET who cast 0.35 votes against their own type and 0.49 votes against Type A players on average (0.14 votes more, or an increase of 40%).

6. Concluding remarks

Ostracism, or exclusion, is often observed in various fields of social and economic life as a particularly strong form of punishment. It has been suggested in the literature that this phenomenon can have desirable effects, by promoting the adherence to social norms and reducing free-riding. At the same time, however, one must not oversee the possibility that a 'dark side' of exclusion also exists, if for instance this severe form of punishment erodes social capital, provokes retaliatory behavior, or systematically disadvantages certain groups of individuals. The aim of this paper has been to offer a comprehensive analysis of these effects, in order to highlight the potential benefits and perils of exclusion. In particular, we have examined endogenous exclusion (i.e., exclusion implemented by majority vote within a team) in a controlled lab experiment where subjects face a social dilemma in terms of a choice between stealing and producing. Additionally, we have introduced inequalities in productivity within teams, in order to investigate how exclusion and reintegration work in an environment that qualitatively imitates real-world economic differences between members of a social group.

Our results show that exclusion has a positive effect on pro-social behavior, by reducing the total amount of theft and leading excluded subjects to steal significantly less after they are re-admitted into their team. However, there are also drawbacks. On the one hand, we find that excluded subjects retaliate against the rest of the team by casting more exclusion votes when they are re-admitted into the team. On the other hand, subjects that have been excluded more frequently in the past are more likely to be sent away from their team again, even after controlling for their theft decisions. This suggests that a successful integration may be particularly problematic for individuals who have been stigmatized by exclusion. In terms of heterogeneity in the teams, we find that economic inequality (in terms of production possibilities) generally increases theft, and that subjects are more likely to steal from and vote against subjects who belong to their out-group. This group bias is not purely driven exclusively by distributional concerns, since we find a similar pattern in stealing and voting in the minimal group treatments.

While we are aware of the inherent limitations in generalizing findings from the economic laboratory, we believe that our result can contribute to the policy debate regarding the effectiveness of correctional systems in promoting rehabilitation and reintegration of inmates into social and economic life. We have shown that, in a controlled lab environment, members who are excluded from a team are less likely to engage in anti-social behavior (operationalized and framed as theft in our experiment) when they return to the team. This is an encouraging result, which points towards a positive correctional effect of exclusion. On the other hand, however, there is evidence in our experiment that exclusion can backfire. We find that returning team members retaliate by making efforts to exclude other team members in the future. What is even more worrying is that returning members seem to carry a stigma with them, in the form of a higher likelihood of being excluded again and again, even after controlling for their activity and decisions during the team interaction. Obviously, the benefits and perils of excluding individuals from a team must be weighed against each other. Doing so is not a goal of this study, since in an experiment this weighing would depend on the chosen parameterization. Our goal has been to assess and highlight the various mechanisms that can create positive and negative effects of exclusion on individual and team outcomes.

Declarations

Funding: The project has been approved and received funding by the vice-rectorate for research of the University of Innsbruck (project number 282227).

Acknowledgements: We are grateful to Marie-Claire Villeval and seminar participants at the University of Düsseldorf, Max Planck Institute for Tax Law and Public Finance, University Paris Nanterre, Monash University and University of Queensland, and participants at the 2019 ESA conference in Dijon for helpful comments and suggestions. We thank Julia Freuding for providing assistance in this research.

Conflicts of interest/Competing interests: The authors have no conflicts of interest no declare.

Availability of data and material: All data are available from the corresponding author upon request.

Code availability: The z-tree codes used for the experiment is available from the corresponding author upon request.

References

Ahn, T. K., Balafoutas, L., Batsaikhan, M., Campos-Ortiz, F., Putterman, L., & Sutter, M. (2016). Securing property rights: A dilemma experiment in Austria, Mexico, Mongolia, South Korea and the United States. *Journal of Public Economics*, *143*, 115–124.

Ahn, T. K., Loukas, B., Batsaikhan, M., Campos-Ortiz, F., Putterman, L., & Sutter, M. (2018). Trust and communication in a property rights dilemma. *Journal of Economic Behavior and Organization*, *149*, 413–433.

Akpalu, W., & Martinsson, P. (2012). Ostracism and common pool resource management in a developing country: Young fishers in the laboratory. *Journal of African Economies*, *21*(2), 266–306.

Balafoutas, L., García-Gallego, A., Georgantzis, N., Jaber-Lopez, T., & Mitrokostas, E. (2020). Rehabilitation and social behavior: Experiments in prison. *Games and Economic Behavior*, *119*, 148–171.

Baumeister, R. F., & Leary, M. R. (1995). The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation. *Psychological Bulletin*, *117*(3), 497–529.

Bayer, P., Hjalmarsson, R., & Pozen, D. (2009). Building criminal capital behind bars: Peer effects in juvenile corrections. *Quarterly Journal of Economics*, *124*(1), 105–147.

Bernstein, M. J., & Claypool, H. M. (2012). Not all social exclusions are created equal: Emotional distress following social exclusion is moderated by exclusion paradigm. *Social Influence*, *7*(2), 113–130.

Bhuller, M., Dahl, G., Løken, K., & Mogstad, M. (2018). *Incarceration spillovers in criminal and family networks (No. w24878)*. National Bureau of Economic Research.

Bhuller, M., Dahl, G., Løken, K., & Mogstad, M. (2020). Incarceration, recidivism and employment. *Journal of Political Economy*, *128*(4), 1269–1324.

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

Buckley, E., & Croson, R. (2006). Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics*, *90*(4–5), 935–955.

Charness, G., & Yang, C. L. (2014). Starting small toward voluntary formation of efficient large groups in public goods provision. *Journal of Economic Behavior and Organization*, *102*, 119–132.

Chen, M. K., & Shapiro, J. M. (2007). Do harsher prison conditions reduce recidivism? A discontinuity-based approach. *American Law and Economics Review*, *9*(1), 1–29.

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, *99*(1), 431–457.

Cinyabuguma, M., Page, T., & Putterman, L. (2005). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics*, *89*, 1421–1435.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, *9*(3), 265–279.

Dannenberg, A., Haita-Falah, C., & Zitzelsberger, S. (2019). Voting on the threat of exclusion in a public goods experiment. *Experimental Economics*, 1–26.

Davis, B. J., & Johnson, D. B. (2015). Water Cooler Ostracism: Social Exclusion as a Punishment Mechanism. *Eastern Economic Journal*, *41*(1), 126–151.

Drago, F., Galbiati, R., & Vertova, P. (2009). The deterrent effects of prison: Evidence from a natural experiment. *Journal of Political Economy*, *117*(2), 257–280.

Durose, M., Cooper, A., & Snyder, H. (2014). *Recidivism of prisoners released in 30 states in 2005 : Patterns from 2005 to 2010*. Washington, DC: US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science*, *25*(3), 656–664.

Fischbacher, U. (2007). Z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Gruter, M., & Masters, R. D. (1986). Ostracism as a social and biological phenomenon: An introduction. *Ethology and Sociobiology*, *7*(3–4), 149–158.

Gürerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*(5770), 108–111.

Güth, W., Levati, M. V., Sutter, M., & van der Heijden, E. (2007). Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics*, *91*(5–6), 1023–1042.

Hermann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367.

Hirshleifer, D., & Rasmusen, E. (1989). Cooperation in a repeated prisoners' dilemma with ostracism. *Journal of Economic Behavior and Organization*, *12*(1), 87–106.

Kingsley, D. C. (2016). Endowment heterogeneity and peer punishment in a public good experiment: Cooperation and normative conflict. *Journal of Behavioral and Experimental Economics*, *60*, 49–61.

Lowen, A., & Schmitt, P. (2013). Cooperation limitations under a one-time threat of expulsion and punishment. *Journal of Socio-Economics*, *44*, 68–74.

Maier-Rigaud, F. P., Martinsson, P., & Staffiero, G. (2010). Ostracism and the provision of a public good: experimental evidence. *Journal of Economic Behavior and Organization*, *73*, 387–395.

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

Masters, R. D. (1984). Ostracism, voice, and exit: the biology of social participation. *Social Science Information*, *23*(6), 877–893.

Neuhofer, S., & Kittel, B. (2015). Long-and short-term exclusion in the public goods game: An experiment on ostracism.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, *92*(1–2), 91–112.

Reuben, E., & Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior*, *77*(1), 122–137.

Sheremeta, R. M., Tucker, S. J., & Zhang, J. (2009). *Creating self-sustained social norms through communication and ostracism*. 19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011.

Solda, A., & Villeval, M. C. (2020). Exclusion and reintegration in a social dilemma. *Economic Inquiry*, *58*(1), 120–149.

Tan, F. (2008). Punishment in a linear public good game with productivity heterogeneity. *De Economist*, *156*(3), 269–293.

van Beest, I., & Williams, K. D. (2006). When inclusion costs and ostracism pays, ostracism still hurts. *Journal of Personality and Social Psychology*, *91*(5), 918–928.

Williams, K. D. (1997). Social ostracism. In *Aversive Interpersonal Behaviors* (pp. 133–170). Springer, Boston, MA.

Williams, K. D. (2002). *Ostracism: The power of silence*. Guilford Press.

Williams, K. D. (2007). Ostracism. *Annual Review of Psychology*, *58*(1), 425–452.

Williams, K. D., Cheung, C. K. T., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the internet. *Journal of Personality and Social Psychology*, *79*(5), 748–762.

Williams, K. D., Govan, C. L., Croker, V., Tynan, D., Cruickshank, M., & Lam, A. (2002). Investigations into differences between social- and cyberostracism. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 65–77.

Williams, K. D., & Zadro, L. (2001). Ostracism: On being ignored, excluded, and rejected. In M. R. Leary (Ed.), *Interpersonal rejection* (pp. 21–53).

Zadro, L., Williams, K. D., & Richardson, R. (2004). How low can you go? Ostracism by a computer is sufficient to lower self-reported levels of belonging, control, self-esteem, and meaningful existence. *Journal of Experimental Social Psychology*, *40*(4), 560–567.

Appendix

A.1. Additional figures

Figure A1: Screen production/ stealing decision

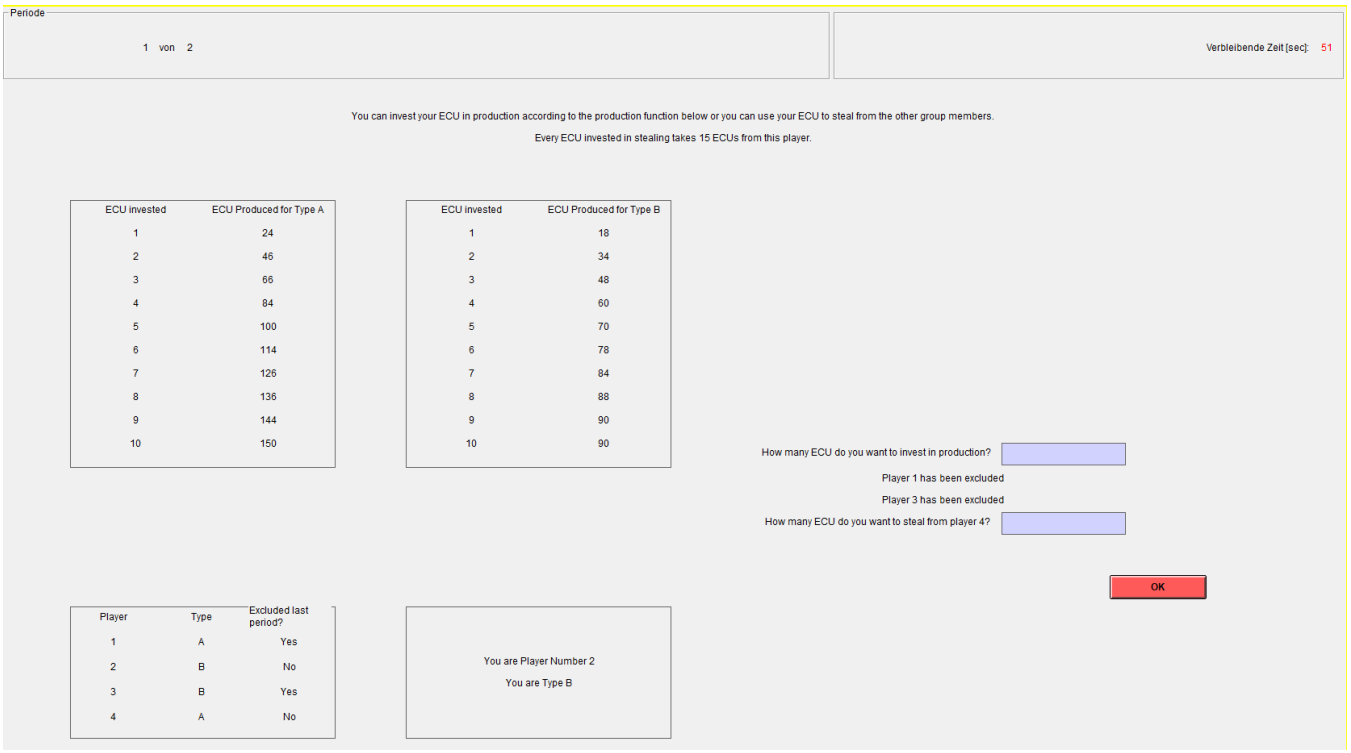


Figure A2: Information table and voting decision

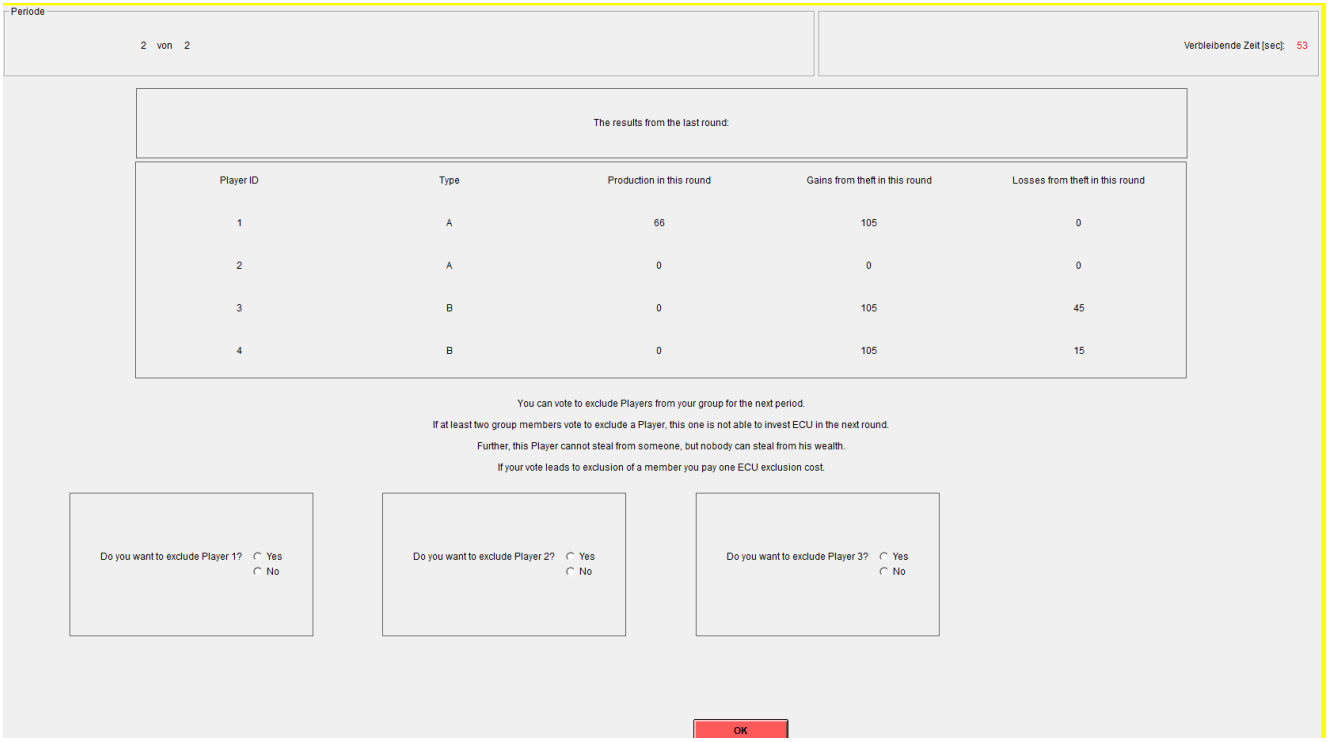


Figure A3: First exclusion decision in *HET*

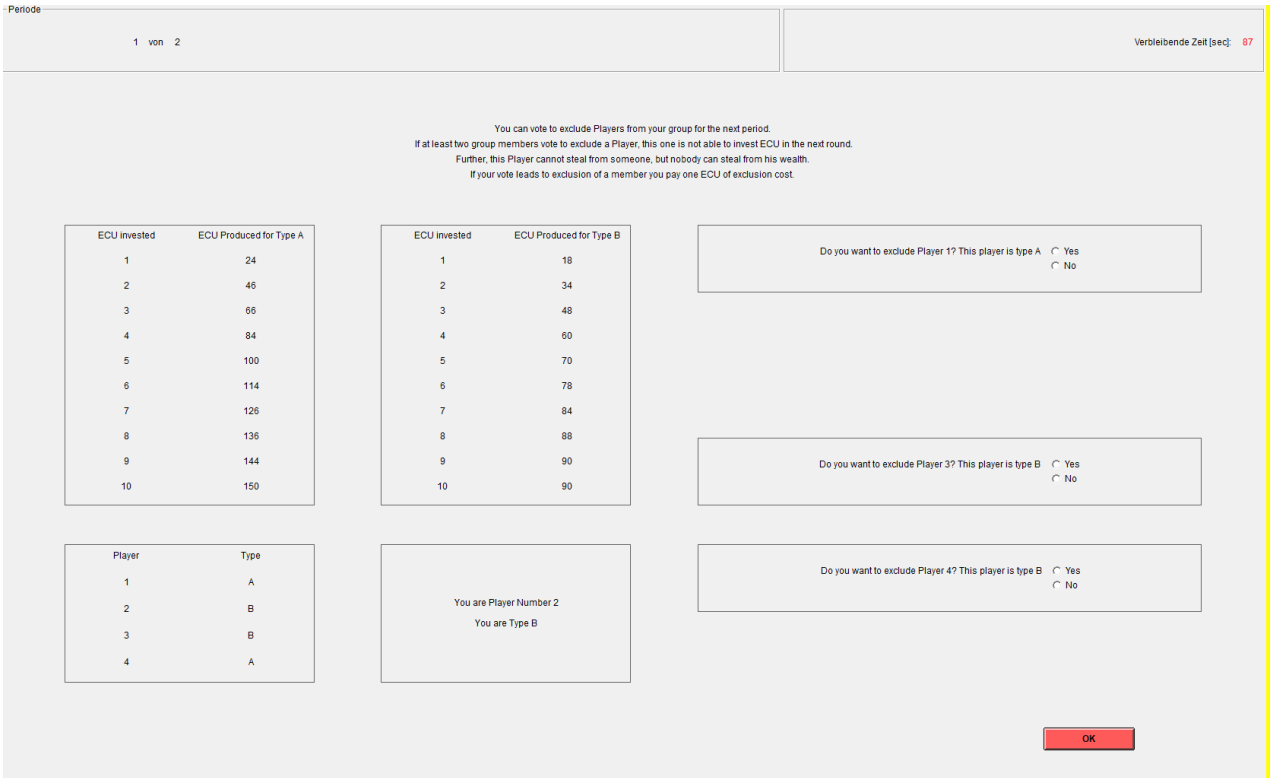


Figure A4: First period stealing (in tokens) in *BASE*

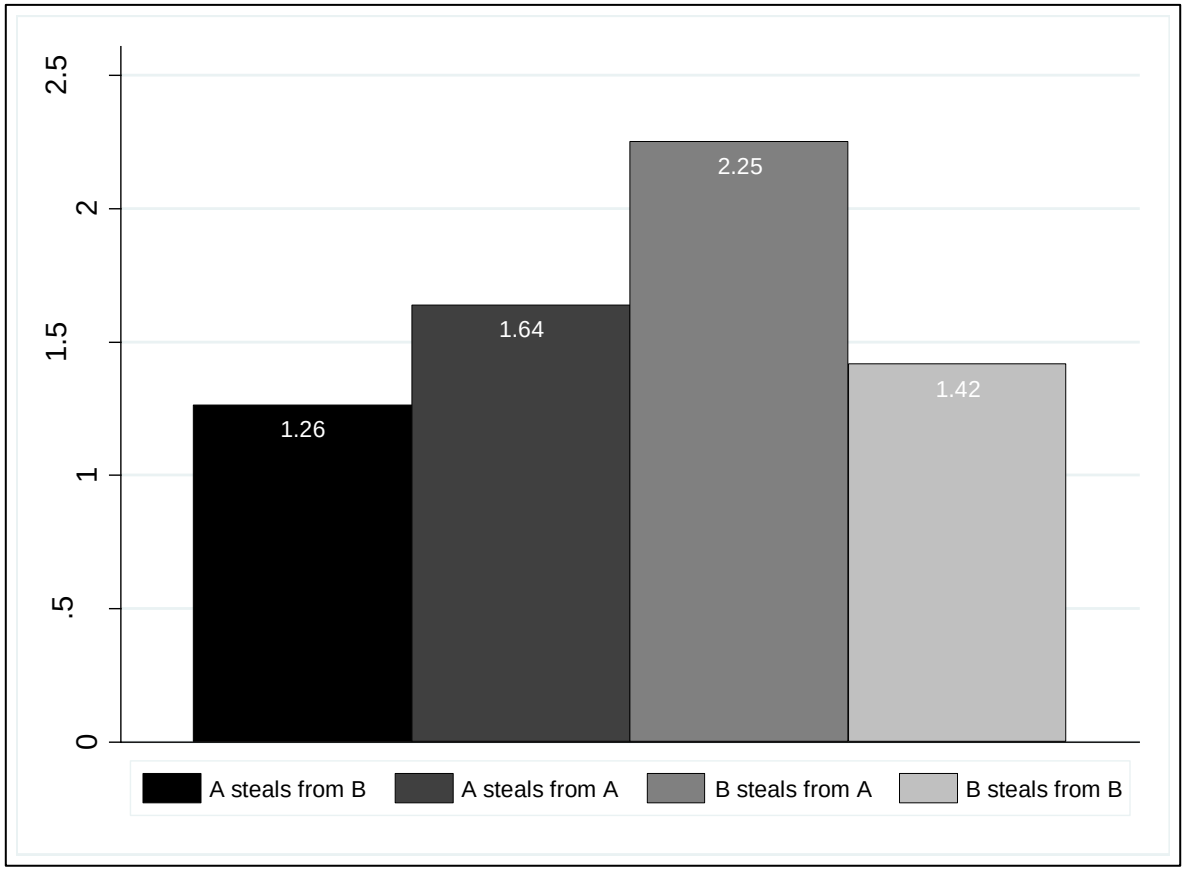
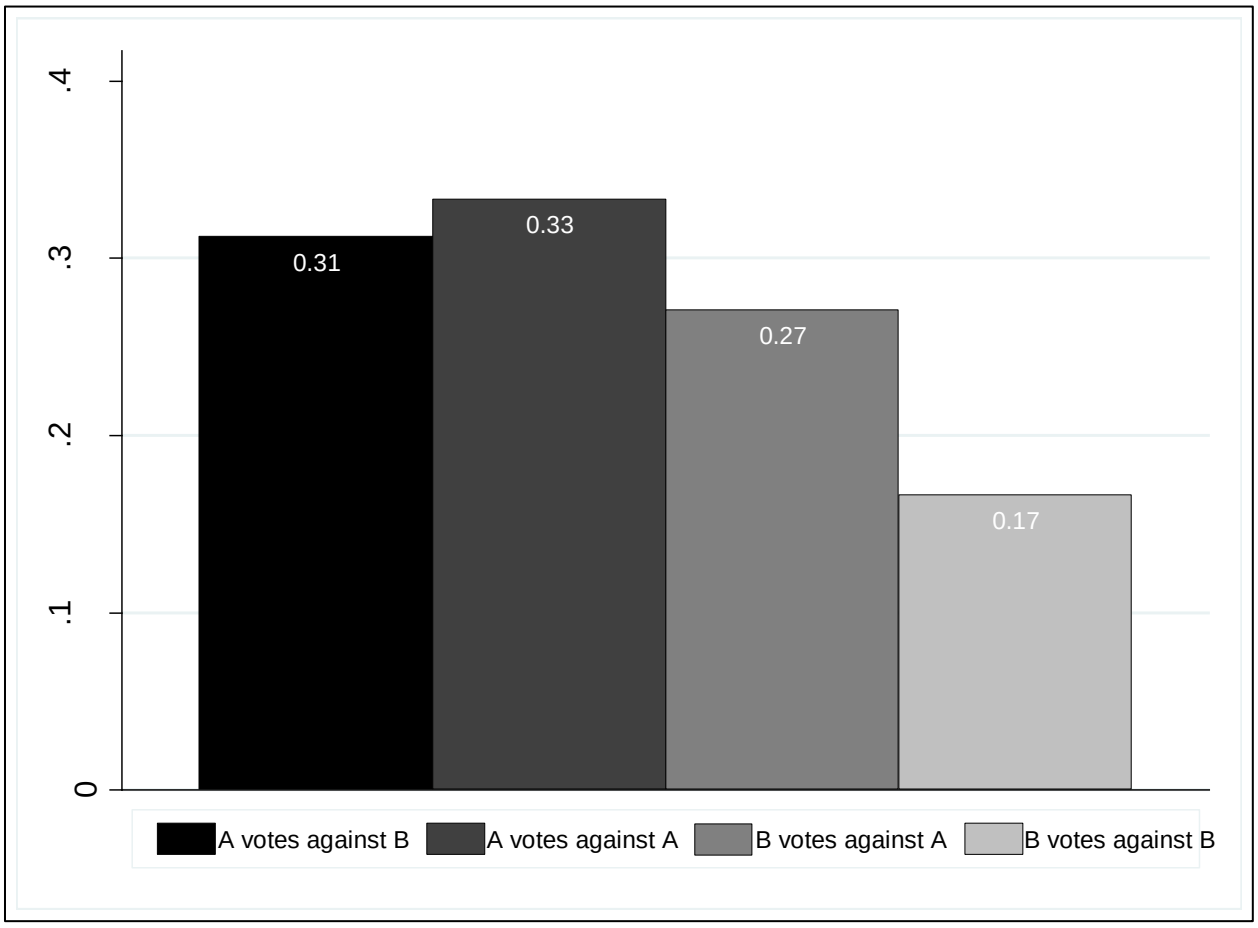


Figure A5: First period voting in *HET*



A.2. Additional tables

Table A1: Determinants of stealing in *HET*, Tobit regressions

VARIABLES	(1) Theft	(2) Theft	(3) $Theft_t$ - $Theft_{t-2}$	(4) $Theft_t$ - $Theft_{t-2}$
TypeA	-0.366**	1.142***	-0.841**	4.367***
	(0.181)	(0.412)	(0.366)	(0.950)
$Exclusion_{t-2}$	-0.504***	-0.463**	1.577***	-1.815**
	(0.186)	(0.227)	(0.566)	(0.704)
$Exclusion_{t-2}$ *x TypeA*		-0.097		0.352
		(0.275)		(0.841)
MeanTheft	1.217***	1.162***	0.217	0.114
	(0.086)	(0.091)	(0.174)	(0.195)
MeanTheft x TypeA		0.178		0.466**
		(0.115)		(0.227)
MeanTheft_Received	0.172***	0.126	0.327**	-0.095
	(0.061)	(0.122)	(0.148)	(0.229)
MeanTheft_Received x TypeA		0.059		0.578***
		(0.139)		(0.222)
Period	0.0732***	0.074***	0.087	0.089
	(0.018)	(0.018)	(0.058)	(0.057)
Team size$_t$	0.711***	0.708***	0.391	0.373
	(0.160)	(0.163)	(0.332)	(0.339)
Female	0.125	0.0973	-0.431	-0.442
	(0.258)	(0.269)	(0.335)	(0.332)
Constant	-4.462***	4.083***	5.336***	3.493***
	(0.572)	(0.556)	(1.281)	(1.288)
Observations	999	999	802	802
Number of teams	24	24	24	24

Notes. Tobit regressions, left-censored at 0. Standard errors are clustered at team level. Dependent variable in (1) and (2): Theft tokens invested by subject *i* in period *t*. Dependent variable in (3) and (4): Theft tokens invested by subject *i* in period *t*, minus tokens invested by subject *i* in period *t-2*. Independent variables described in text. *, **, *** indicates significance at the 10%, 5%, 1% level, respectively.

Table A2: Evolution of stealing, split by high and low theft teams

VARIABLES	(1) $Theft_t$ - $Theft_{t-2}$ (Below median theft)	(2) $Theft_t$ - $Theft_{t-2}$ (Above median theft)
TypeA	-0.167	-0.152
	(0.250)	(0.264)
$Exclusion_{t-2}$	-1.577***	-0.893***
	(0.305)	(0.256)
MeanTheft	0.136*	-0.044
	(0.079)	(0.066)
MeanTheft_Received	0.049	0.017
	(0.088)	(0.078)
Period	0.141***	0.035
	(0.030)	(0.031)
$Team\ size_t$	0.201	0.584***
	(0.154)	(0.184)
Female	-0.178	-0.058
	(0.228)	(0.272)
Constant	-1.521**	-1.631**
	(0.666)	(0.831)
Observations	399	403
Number of teams	12	12

Notes. Multilevel regressions, with subject and team random effects. Standard errors are clustered at team level. Dependent variable in (1): Theft tokens invested by subject *i* in period *t*, minus tokens invested by subject *i* in period *t-2* for low theft teams. Dependent variable in (2): Theft tokens invested by subject *i* in period *t*, minus tokens invested by subject *i* in period *t-2* for high theft teams. Independent variables described in text. *, **, *** indicates significance at the 10%, 5%, 1% level, respectively.

Table A3: Evolution of stealing in *HOA* and *HOB*

	(1)	(2)
VARIABLES	$Theft_t$ - $Theft_{t-2}$	$Theft_t$ - $Theft_{t-2}$
Exclusion$_{t-2}$	-1.368***	-1.250***
	(0.183)	(0.229)
MeanTheft	0.005	0.009
	(0.010)	(0.014)
MeanTheft_Received	0.017	-0.005
	(0.010)	(0.014)
Period	0.061***	0.151***
	(0.020)	(0.030)
Team size$_t$	0.485***	0.380**
	(0.127)	(0.151)
Female	0.013	-0.022
	(0.162)	(0.200)
Constant	-2.272***	-2.152***
	(0.504)	(0.588)
Observations	634	604
Number of teams	17	17

Notes. Multilevel regressions, with subject and team random effects. Standard errors are clustered at team level. Dependent variable in (1): Theft tokens invested by subject *i* in period *t*, minus tokens invested by subject *i* in period *t-2* for *HOA* treatment. Dependent variable in (2): Theft tokens invested by subject *i* in period *t*, minus tokens invested by subject *i* in period *t-2* for *HOB* treatment. Independent variables described in text. *, **, *** indicates significance at the 10%, 5%, 1% level, respectively.

Table A4: Determinants of voting in *HET*

VARIABLES	(1) Votes Given	(2) Votes Given	(3) Votes Received	(4) Votes Received
TypeA	-0.013	0.089	0.024	0.047
	(0.044)	(0.086)	(0.021)	(0.046)
$Exclusion_{t-2}$	0.087***	0.141***	-0.030	-0.038
	(0.025)	(0.035)	(0.023)	(0.032)
$Exclusion_{t-2}$ *x TypeA*		-0.111**		0.014
		(0.050)		(0.044)
MeanTheft	0.005	0.010	0.020***	0.017**
	(0.009)	(0.012)	(0.005)	(0.007)
MeanTheft x TypeA		-0.016		0.007
		(0.019)		(0.011)
MeanTheft_Received	0.026**	0.028*	-0.004	-0.000
	(0.010)	(0.016)	(0.006)	(0.010)
MeanTheft_Received x TypeA		-0.002		-0.007
		(0.020)		(0.012)
Cumulative_Exclusions			0.135***	0.142***
			(0.008)	(0.010)
Cumulative_Exclusions x TypeA				-0.014
				(0.012)
Period	0.000	0.000	-0.030***	-0.030***
	(0.002)	(0.002)	(0.003)	(0.003)
Female	0.091**	-0.086*	0.029	0.028
	(0.04)	(0.046)	(0.022)	(0.022)
Constant	0.382***	0.331***	0.362***	0.351***
	(0.064)	(0.074)	(0.037)	(0.043)
Observations	964	964	964	964
Number of teams	24	24	24	24

Notes. Multilevel regressions, with subject and team random effects. Standard errors are clustered at team level. Dependent variable in (1) and (2): Votes cast by subject *i* in period *t* defined as the proportional share of all possible votes according to the number of active player in the team. Dependent variable in (3) and (4): Proportional number of possible votes received by subject *i* in period *t*. Independent variables described in text. *, **, *** indicates significance at the 10%, 5%, 1% level, respectively.

A.3 Predictions regarding voting

In order to predict whether a member would vote to exclude another member, we need to compare two things: his or her benefit from excluding the other member, and his or her cost from doing so. These costs and benefits always refer to the following period, in which possible exclusions are carried out (remember that the length of exclusion in the experiment is exactly one period). We abstract from learning effects over time and present calculations based on only the period when voting place and the period that follows. All calculations are based on risk neutrality and on the assumption that all agents are selfish and have no motivation other than maximizing their own payoffs. Hence, we do not consider in-group favoritism, retaliatory or reciprocal considerations, or other related behavioral motivations. This means that subjects will randomly allocate their theft tokens among the members that are currently in the group. Similarly, subjects' votes depend on expectations about theft by other members. We further assume that stealing decisions follow the predictions described in section 3.1 (Type A subjects invest 5 tokens and Type B subjects invest 8 tokens into theft).

A.1. Predictions for the heterogeneous treatment (*HET*)

The calculations are based on a cost-benefit analysis from the perspective of a given team member *i* who must decide whether to vote to exclude another team member (or members). The **benefit** to member *i* from excluding another member (or members) refers to the fact that the excluded member(s) cannot steal in the period that follows the exclusion decision. This leads to a *reduction in expected theft* targeted at member *i*. The size of this benefit depends on the other member's type and on how many members there are in the team. The **cost** to member *i* from excluding another member (or members) refers to the fact that theft by the remaining team members is more likely to be targeted at *i*. This leads to an *increase in expected theft* targeted at *i*. The size of this cost depends on *i*'s type, on the other member's type and on how many members there are in the team. Hence, *i*'s decision on whether to vote against each other member in the team will depend on the net benefit, which is the increase or decrease in expected theft targeted at *i* as a result of excluding another member (or members).[20] We begin by calculating the net benefit from voting to exclude one team member, and later also consider voting to exclude more than one member.

(I) Decision of a **Type B** member.

We first consider the case where member *i* is of Type B and decides whether to vote against the other Type B member. For this, we compare the expected theft targeted at member *i* under strategies "B excludes B" and "B excludes no one".[21] Under "B excludes B", the expected theft against *i* in the next period is (5+5)/2, while under "B excludes no one" it is (5+5+8)/3. Hence, the net increase in theft is:

[20] Note that the voting cost of 1 ECU in case of a successful exclusion must be subtracted from this net benefit, but we omit it from the following calculations in the sake of simplicity because it is never decisive.

[21] All strategies in this section refer to the case of member *i* voting to exclude another member (or members), *and* this other member being indeed excluded. We call these strategies "B excludes B", "B excludes A" etc. for succinctness.

$$\Delta Loss_i = \left(\frac{total\ theft\ tokens\ if\ 1B\ is\ exluded}{number\ of\ subjects\ -1} - \frac{total\ theft\ tokens\ if\ no\ one\ is\ exluded}{number\ of\ subjects}\right)$$
$$* stolen\ amount\ per\ token = \left(\frac{10}{2} - \frac{18}{3}\right) * 15 = -15$$

The resulting $\Delta Loss_i = -15$ indicates that the expected loss to theft for member *i* is lower under strategy "B excludes B" than under "B excludes no one". Intuitively, this is because the decrease in total theft (from 18 to 10) outweighs the increase in the likelihood of theft being targeted at member *i* (from 1/3 to 1/2). Therefore, Type B members have an interest in excluding each other.

We next consider the case where member *i* is of Type B and decides whether to vote against one Type A member. Using the same procedure as above, we compare the expected theft targeted at member *i* under strategies "B excludes 1A" (i.e., B excludes one Type A member) and "B excludes no one". The net increase in theft targeted at *i* is given by:

$$\Delta Loss_i = \left(\frac{total\ theft\ tokens\ if\ 1A\ is\ exluded}{number\ of\ subjects\ -1} - \frac{total\ theft\ tokens\ if\ no\ one\ is\ exluded}{number\ of\ subjects}\right)$$
$$* stolen\ amount\ per\ token = \left(\frac{13}{2} - \frac{18}{3}\right) * 15 = 7.5$$

The resulting $\Delta Loss_i = 7.5$ indicates that the expected theft targeted at a member *i* of Type B is higher if *i* successfully excludes one member of Type A, than if no Type A is excluded. Therefore, Type B members have no interest in excluding a Type A member.

(II) Decision of a **Type A** member.

We first consider the case where member *i* is of Type A and decides whether to vote against one Type B member. We compare the expected theft targeted at member *i* under strategies "A excludes 1B" (i.e., A excludes one Type B member) and "A excludes no one". The net increase in theft targeted at *i* is given by:

$$\Delta Loss_i = \left(\frac{total\ theft\ tokens\ if\ 1B\ is\ exluded}{number\ of\ subjects\ -1} - \frac{total\ theft\ tokens\ if\ no\ one\ is\ exluded}{number\ of\ subjects}\right)$$
$$* stolen\ amount\ per\ token = \left(\frac{13}{2} - \frac{21}{3}\right) * 15 = -7.5$$

Since $\Delta Loss_i$ is negative, we conclude that Type A members have an interest in excluding a Type B member.

Next, we consider the case where member *i* is of Type A and decides whether to vote against the other Type A member. Using the same procedure as above, we compare the expected theft targeted at member *i* under strategies "A excludes A" and "A excludes no one". The net increase in theft targeted at *i* is given by:

$$\Delta Loss_i = \left(\frac{total\ theft\ tokens\ if\ 1A\ is\ exluded}{number\ of\ subjects\ -1} - \frac{total\ theft\ tokens\ if\ no\ one\ is\ exluded}{number\ of\ subjects}\right)$$
$$* stolen\ amount\ per\ token = \left(\frac{16}{2} - \frac{21}{3}\right) * 15 = 15\ ECU$$

Since $\Delta Loss_i$ is positive, we conclude that Type A members have no interest in excluding each other.

(III) Excluding more than one member.

The above analysis examines cases in which each player excludes only one other player. However, other strategies are possible. We calculate below the net increase in expected theft for each of these strategies, always compared to the strategy of not excluding any member.

Type B players:

- Strategy **"Exclude 1B+1A"** player: The net increase in expected theft targeted at *i* is $\left(5 - \frac{18}{3}\right) * 15 = -15$ (because there is only one A player left, who will steal 5 ECU in the next period).
- Strategy **"Exclude 1B+2A"** players: In this case, the only member left in the team is member *i* and no theft takes place. Hence, instead of the equilibrium payoff of 64 (see footnote 9), *i* receives a payoff of 90 from investing everything into production. This is the best possible scenario from *i*'s perspective, since the social dilemma does not exist.
- Strategy **"Exclude 2A"** players: The net increase in expected theft targeted at *i* is $\left(8 - \frac{18}{3}\right) * 15 = 30$ (because there is only one B player left, who will steal 8 ECU in the next period).

We can now compare the six possible strategies of a Type B player discussed so far, in terms of associated payoffs: **Exclude (2B+1A) > Exclude (1B+1A) = Exclude (1B) > Exclude no one > Exclude (1A) > Exclude (2A).**

Type A players:

- Strategy **"Exclude 1B+1A"** players: The net increase in expected theft targeted at *i* is $\left(8 - \frac{18}{3}\right) * 15 = 30$ (because there is only one A player left, who will steal 5 ECU in the next period).
- Strategy **"Exclude 1A+2B"** players: In this case, the only member left in the team is member *i* and no theft takes place. Hence, instead of the equilibrium payoff of 70 (see footnote 9), *i* receives a payoff of 150 from investing everything into production. This is the best possible scenario from *i*'s perspective.
- Strategy **"Exclude 2B"** players: The net increase in expected theft targeted at *i* is $\left(5 - \frac{18}{3}\right) * 15 = -15$ (because there is only one B player left, who will steal 8 ECU in the next period).

We can now compare the six possible strategies of a Type A player discussed so far, in terms of associated payoffs: **Exclude (2B+1A) > Exclude (2B) > Exclude (1B) > Exclude no one > Exclude (1A) = Exclude (1A+1B)**

The above calculations reveal that excluding all players yields the highest benefit for any individual player, but this strategy cannot be part of a pure strategy equilibrium since it leads to everyone being excluded (and paying the voting cost without gaining anything). The second highest net benefit,

both to Type A and to Type B players, comes from voting to exclude both type B players. Type B players thus anticipate that they will be excluded and do not vote, in order to save the voting costs.

We summarize our equilibrium predictions for voting as follows. All members (regardless of their type) have an interest in excluding low productivity (Type B) members but not in excluding high productivity (Type A) members. However, the net benefits calculated above apply only to team members who will be in the team in the following period and hence will have a stake in the game. If a member expects to be excluded, then she has nothing to gain by excluding another member and she has no reason to pay the voting cost. Hence, we can expect the following voting behavior in the pure strategy equilibrium:

(1) Type B members do not cast any vote. If they did (for instance, by voting against each other), they would have to pay the cost of 1 ECU, but they would not realize net benefit calculated since in the following period they will be excluded. This assumes that they correctly anticipate the strategy of Type A members described in (2).

(2) Correctly anticipating the fact that Type B members do not cast any vote, both Type A members vote to exclude the two Type B members since this increases their payoff (by reducing expected theft). Both Type A members are pivotal in this case and must cast two votes each.

(3) In addition to excluding the two Type B members, each Type A member has an interest in excluding the other member of the same type as well, since in that case they would be alone in the group and invest everything into production. However, given (1), no majority will be reached in which a Type A member is excluded. Hence, there is no reason for Type A members to vote against each other and they are both indifferent in this respect.

Strategies (1) and (2) represent mutual best responses and form a pure strategy Nash equilibrium.

A.2. Predictions for the homogeneous treatments (HOA and HOB)

Replicating the above calculations for teams consisting of four Type A or four Type B members, it is straightforward to show that the net benefit to member *i* from excluding any member *j* is zero, regardless of the actual number of members in the team. This is due to the fact that now all members are the same, and what *i* loses from theft in expectation she also gains from theft in expectation, from any other member *j*. In this case, the indifference is broken through the introduction of the small voting cost, hence our prediction here is that no votes are cast to exclude other members in any of the two homogeneous treatments.

Online Appendix (for online publication only)

Instructions for treatment *HET* (translated from German):

Dear participants,

Welcome to this experiment and thank you for your participation.

Please read the instructions carefully. Your payment depends on the decisions you make, as well as the decisions of the other participants. If you have any questions during the experiment, please raise your hand. Your question will be answered privately. After the experiment, you will be paid privately and in cash. **All experimental payoffs are calculated in Experimental Currency Units (ECU), and will be converted into cash at the exchange rate of 80 ECU = 1€.**

After your departure from the lab, your decision data and any subsequent analysis will be anonymously stored.

Please do not talk to the other participants from now on and do not use any other aid than those provided in this experiment.

Instructions

This experiment consists of 15 rounds. You will be randomly allocated in a group of four subjects and you will remain in the same group in all rounds. You will be randomly assigned for the whole experiment one of two types: either Type A or Type B.

In each round you get an endowment of 10 ECU. All 10 ECU must be allocated, either to production or to stealing from other group members. Your earnings will depend on how much you produce, how much you steal, and how much other group members steal from you, as we describe in detail below.

Production:

The production functions for Type A and Type B are given in the table below and will not change over the game.

For example, Type A group members receive 24 ECU for the first token invested in production, 22 additional ECU for the second token, 20 ECU for the third token, and so on, up to 6 ECU for the last invested token.

Type B group members receive 18 ECU for the first token invested in production, 16 additional ECU for the second token, 14 ECU for the third token, and so on, up to 0 ECU for the last token.

Tokens invested	Total ECU produced Type A	Total ECU produced Type B
1	24	18
2	46	34
3	66	48
4	84	60
5	100	70
6	114	78
7	126	84
8	136	88
9	144	90
10	150	90

Stealing:

For each token you invest in stealing from a given group member, you receive 15 ECU from that member. Equally, if another group member invests tokens to steal from you, you will lose 15 ECU from your wealth for each token invested.

Exclusion:

In every round you have the opportunity to vote in order to exclude other group members. The exclusion lasts for one round. This exclusion depends on your vote and the votes from the other group members. Excluded subjects do not participate in the group, i.e., they are not able to produce or steal anything, they cannot vote, and nobody can steal from them. During the round of exclusion they receive a payment of 5 ECU.

See below the details on how the voting works:

- If in a given period there are four members in the group (i.e., no-one was excluded in the previous period), then each member casts three votes, indicating whether she wants to exclude each of the other three members (you cannot vote for yourself). If at least two votes are cast to exclude a given group member, this member is excluded for the following period.

- If in a given period there are three members in the group (i.e., one member was excluded in the previous period), then each member casts two votes, indicating whether she wants to exclude each of the other two members (you cannot vote for yourself). If two votes suggest to exclude a given group member, this member is excluded for the following period.

- If in a given period there are two members in the group (i.e., two members were excluded in the previous period), then there is no voting for exclusion.

Notice that, if you voted to exclude another member and the exclusion takes place, you will have to pay 1 ECU from your wealth. If you voted to exclude a player but the exclusion does not take place, then your vote remains costless.

You will be asked to indicate your vote in each round for each group member by ticking the corresponding box on your screen:

Timing of decisions:

The exclusion vote will take place at the end of each period, after every member has been informed about the token allocations of all other group members. Hence, the timing in each period will be as follows:

1. Members allocate their 10 tokens between production and theft.

2. Members are informed about each other's allocations and the exclusion vote takes place.

An exception is period 1, which includes an initial exclusion vote before allocation or theft. From period 2 onwards, the timing is as described above.

Earnings:

At the beginning of period 1, you receive an **initial endowment of 200 ECU**. From period 2 onwards, your earnings in each period are calculated in the following way:

Earnings= [Benefit from production] + [Benefit from stealing] – [Loss from theft] – [Cost from voting to exclude]

The total earnings from the experiment are the **sum of your earnings over all 15 rounds**.

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

https://www.uibk.ac.at/eeecon/wopec/

2021-19 **Alexandra Baier, Loukas Balafoutas, Tarek Jaber-Lopez:** Ostracism and Theft in Heterogeneous Groups

2021-18 **Zvonimir Bašić, Parampreet C. Bindra, Daniela Glätzle-Rützler, Angelo Romano, Matthias Sutter, Claudia Zoller:** The roots of cooperation

2021-17 **Silvia Angerer, Jana Bolvashenkova, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Children's patience and school-track choices several years later: Linking experimental and field data

2021-16 **Daniel Gründler, Eric Mayer, Johann Scharler:** Monetary Policy Announcements, Information Schocks, and Exchange Rate Dynamics

2021-15 **Sebastian Bachler, Felix Holzmeister, Michael Razen, Matthias Stefan:** The Impact of Presentation Format and Choice Architecture on Portfolio Allocations: Experimental Evidence

2021-14 **Jeppe Christoffersen, Felix Holzmeister, Thomas Plenborg:** What is Risk to Managers?

2021-13 **Silvia Angerer, Daniela Glätzle-Rützler, Christian Waibel:** Trust in health care credence goods: Experimental evidence on framing andsubject pool effects

2021-12 **Rene Schwaiger, Laura Hueber:** Do MTurkers Exhibit Myopic Loss Aversion?

2021-11 **Felix Holzmeister, Christoph Huber, Stefan Palan:** A Critical Perspective on the Conceptualization of Risk in Behavioral and Experimental Finance

2021-10 **Michael Razen, Alexander Kupfer:** Can increased tax transparency curb corporate tax avoidance?

2021-09 **Changxia Ke, Florian Morath, Anthony Newell, Lionel Page:** Too big to prevail: The paradox of power in coalition formation

2021-08 **Marco Haan, Pim Heijnen, Martin Obradovits:** Competition with List Prices

2021-07 **Martin Dufwenberg, Olof Johansson-Stenman, Michael Kirchler, Florian Lindner, Rene Schwaiger:** Mean Markets or Kind Commerce?

2021-06 **Christoph Huber, Jürgen Huber, and Michael Kirchler:** Volatility Shocks and Investment Behavior

2021-05 **Max Breitenlechner, Georgios Georgiadis, Ben Schumann:** What goes around comes around: How large are spillbacks from US monetary policy?

2021-04 **Utz Weitzel, Michael Kirchler:** The Banker's Oath And Financial Advice

2021-03 **Martin Holmen, Felix Holzmeister, Michael Kirchler, Matthias Stefan, Erik Wengström:** Economic Preferences and Personality Traits Among Finance Professionals and the General Population

2021-02 **Christian König-Kersting:** On the Robustness of Social Norm Elicitation

2021-01 **Laura Hueber, Rene Schwaiger:** Debiasing Through Experience Sampling: The Case of Myopic Loss Aversion.

2020-34 **Kai A. Konrad, Florian Morath:** The Volunteer's Dilemma in Finite Populations

2020-33 **Katharina Momsen, Markus Ohndorf:** Expressive Voting vs. Self-Serving Ignorance

2020-32 **Silvia Angerer, Daniela Glätzle-Rützler, Christian Waibel:** Monitoring institutions in health care markets: Experimental evidence

2020-31 **Jana Friedrichsen, Katharina Momsen, Stefano Piasenti:** Ignorance, Intention and Stochastic Outcomes

2020-30 **Esther Blanco, Alexandra Baier, Felix Holzmeister, Tarek Jaber-Lopez, Natalie Struwe:** Substitution of social concerns under the Covid-19 pandemic

2020-29 **Andreas Hackethal, Michael Kirchler, Christine Laudenbach, Michael Razen, Annika Weber:** On the (ir)relevance of monetary incentives in risk preference elicitation experiments

2020-28 **Andrej Gill, Matthias Heinz, Heiner Schumacher, Matthias Sutter:** Trustworthiness in the Financial Industry

2020-27 **Matthias Sutter, Michael Weyland, Anna Untertrifaller, Manuel Froitzheim:** Financial literacy, risk and time preferences - Results from a randomized educational intervention

2020-26 **Rene Schwaiger, Jürgen Huber, Michael Kirchler, Daniel Kleinlercher, Utz Weitzel:** Unequal Opportunities, Social Groups, and Redistribution

2020-25 **Roman Inderst, Martin Obradovits:** Competitive Strategies when Consumers are Relative Thinkers: Implications for Pricing, Promotions, and Product Choice

2020-24 **Martin Obradovits, Philipp Plaickner:** Price-Directed Search and Collusion

2020-23 **Helena Fornwagner, Oliver P. Hauser:** Climate action for (my) children

2020-22 **Esther Blanco, Natalie Struwe, James M. Walker:** Incentivizing public good provision through outsider transfers: experimental evidence on sharing rules and additionality requirements

2020-21 **Loukas Balafoutas, Helena Fornwagner, Rudolf Kerschbamer, Matthias Sutter, Maryna Tverdostup:** Diagnostic Uncertainty and Insurance Coverage in Credence Goods Markets

2020-20 **Anna Ulrichshofer, Markus Walzl:** Customer Disputes, Misconduct, and Reputation Building in the Market for Financial Advice

2020-19 **Anna Ulrichshofer, Markus Walzl:** Social Comparison and Optimal Contracts in the Competition for Managerial Talent

2020-18 **Martin Obradovits, Philipp Plaickner:** Searching for Treatment

2020-17 **Jun Honda:** The Gender-Punishment Gap revisited

2020-16 **Jun Honda:** The Relation between Rankings and Risk-Taking in the Labor Market for Financial Advice

2020-15 **Christina Bannier, Eberhard Feess,Natalie Packham, Markus Walzl:** Differentiation and Risk-Aversion in Imperfectly Competitive Labor Markets

2020-14 **Felix Holzmeister, Rudolf Kerschbamer:** oTree: The Equality Equivalence Test

2020-13 **Parampreet Christopher Bindra, Graeme Pearce:** The effect of priming on fraud: Evidence from a natural field experiment

2020-12 **Alessandro De Chiara, Marco A. Schwarz:** A Dynamic Theory of Regulatory Capture

2020-11 **Christoph Huber, Jürgen Huber, Michael Kirchler:** Market shocks and professionals' investment behavior - Evidence from the COVID-19 crash

2020-10 **Elisabeth Gsottbauer, Daniel Müller, Samuel Müller, Stefan T. Trautmann, Galina Zudenkova:** Social class and (un)ethical behavior: Causal versus correlational evidence

2020-09 **Parampreet Christopher Bindra, Rudolf Kerschbamer, Daniel Neururer, Matthias Sutter:** Reveal it or conceal it: On the value of second opinions in a low-entry-barriers credence goods market

2020-08 **Robert Steiger, Eva Posch, Gottfried Tappeiner, Janette Walde:** Effects of climate change on tourism demand considering individual seasonal preferences

2020-07 **Fang Liu, Alexander Rasch, Marco A. Schwarz, Christian Waibel:** The role of diagnostic ability in markets for expert services

2020-06 **Matthias Stefan, Jürgen Huber, Michael Kirchler, Matthias Sutter, Markus Walzl:** Monetary and Social Incentives in Multi-Tasking: The Ranking Substitution Effect

2020-05 **Michael Razen, Jürgen Huber, Laura Hueber, Michael Kirchler, Matthias Stefan:** Financial Literacy, Economic Preferences, and Adolescents' Field Behavior

2020-04 **Christian König-Kersting, Johannes Lohse, Anna Louisa Merkel:** Active and Passive Risk-Taking

2020-03 **Christoph Huber, Jürgen Huber:** Bad bankers no more? Truth-telling and (dis)honesty in the finance industry

2020-02 **Dietmar Fehr, Daniel Müller, Marcel Preuss:** Social Mobility Perceptions and Inequality Acceptance

2020-01 **Loukas Balafoutas, Rudolf Kerschbamer:** Credence goods in the literature: What the past fifteen years have taught us about fraud, incentives, and the role of institutions

University of Innsbruck

Working Papers in Economics and Statistics

2021-19

Alexandra Baier, Loukas Balafoutas, Tarek Jaber-Lopez

Ostracism and Theft in Heterogeneous Groups

Abstract
Ostracism, or exclusion by peers, has been practiced since ancient times as a severe form of punishment against transgressors of laws or social norms. The purpose of this paper is to offer a comprehensive analysis on how ostracism affects behavior and the functioning of a social group. We present data from a laboratory experiment, in which participants face a social dilemma on how to allocate limited resources between a productive activity and theft, and are given the opportunity to exclude members of their group by means of majority voting. Our main treatment features an environment with heterogeneity in productivity within groups, thus creating inequalities in economic opportunities and income. We find that exclusion is an effective form of punishment and decreases theft by excluded members once they are re-admitted into the group. However, it also leads to some retaliation by low-productivity members. A particularly worrisome aspect of exclusion is that punished group members are stigmatized and have a higher probability of facing exclusion again. We discuss implications of our findings for penal systems and their capacity to rehabilitate prisoners.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)