Is it a Norm to Favour Your Own Group?*

Donna Harris[1] Benedikt Herrmann[2]

Andreas Kontoleon[3] Jonathan Newton[4]

June, 2014

Abstract

This paper examines the relationship between norm enforcement and in-group favouritism behaviour. Using a new two-stage allocation experiment with punishments, we investigate whether in-group favouritism is considered as a social norm in itself or as a violation of a different norm, such as egalitarian norm. We find that which norm of behaviour is enforced depends on *who the punisher is*. If the punishers belong to the in-group, in-group favouritism is considered a norm and it does not get punished. If the punishers belong to the out-group, in-group favouritism is frequently punished. If the punishers belong to no group and merely observe in-group favouritism (the third-party), they do not seem to care sufficiently to be willing to punish this behaviour. Our results shed a new light on the effectiveness of altruistic norm enforcement when group identities are taken into account and help to explain why in-group favouritism is widespread across societies.

*JEL: C92, D70, D73.*Keywords: In-group Favouritism, Group Identity, Social Norms, In-group Punishment, Out-group Punishment, Third-party Punishment.

*We are grateful for the financial assistance from the Economic and Social Research Council (ESRC) Grant Number: RG58935; the British Academy; the Leverhulme Trust; and the Suzy Paine Trust of the Faculty of Economics, University of Cambridge. We are also grateful for the discussions with Gary Charness, Vincent Crawford, Simon Gaechter, Daniele Nosenzo, Jan Potters, Aldo Rustichini, and Tim Salmon.

[1] **Corresponding author**: Department of Economics, University of Oxford, Manor Road, Oxford, OX1 3UQ/ Somerville College, Woodstock Road, Oxford, OX2 6HD.Email:donhatai.harris@economics.ox.ac.uk.

[2] Behavioural Economics Team, Institute for Health and Consumer Protection, Joint Research Centre, European Commission.

[3] Environmental Economy and Policy Research Group, Department of Land Economy, University of Cambridge, 19 Silver Street, Cambridge, UK, CB3 9EP.

[4] School of Economics, University of Sydney, Sydney, NSW 2006, Australia. The author is a recipient of a Discovery Early Career Researcher Award funded by the Australian Research Council.

1 Introduction

In-group favouritism is preferential treatment given to in-group member(s) at the expense of outsiders (Becker, 1957). It is often perceived as unfair and unacceptable behaviour because it violates the principle of meritocratic competition for resources and economic opportunities. On a larger scale, it can lead to inefficiency, income inequality, and even economic crisis (Prendergast and Topel, 1996; Barr and Oduro, 2002; Fisman, 2003; Bandiera et al., 2009; Anderson, 2010; Mitsopoulos et al., 2011). However, despite the negative perception and detrimental economic and social consequences of this behaviour, it is still widespread around the world (Global Competitiveness Report, 2013).

The literature on group identity and in-group favouritism is vast (e.g. Tajfel et al., 1971; Tajfel and Turner, 1986; Bernhard et al., 2006a; b; Goette et al., 2006; Charness et al., 2007; McLeish and Oxoby, 2007; Guth et al., 2008; Belot and van de Ven, 2011; for a review, see Chen and Li, 2009). One of the main results from these studies is that, given the opportunity, most people favoured their own group, regardless of how the groups were formed. This seemingly natural tendency to favour one's own group stands in sharp contrast to negative perceptions of this behaviour. It also raises a question about whether the normative statement that in-group favouritism is unfair and unacceptable translates into an active enforcement of a social norm[5]. The main contribution of this paper is to address this question by examining whether in-group favouritism is considered a norm in itself or is considered a violation of a different norm, such as egalitarian norm.

This paper defines social norms as behavioural standards based on socially shared beliefs about how individuals ought to behave in a given situation (Bernhard et al., 2006a;b). The need for such a social norm arises when an action generates negative externalities that affect other members of the same society (Fehr and Fischbacher, 2004). But social norms can only be sustained if they are enforced, typically by people's willingness to impose sanctions on the norm violators even at a cost (Fehr and Gachter, 2000; Fehr and Fischbacher, 2004; Hoff et al., 2011; Schram and Charness, 2013). According to this literature, a social norm exists and is sustained only if its enforcement is credible and effective in inducing behavioural change.

There have been a few studies on social norms and in-group favouritism (Bernhard et al., 2006a; b; Goette et al., 2006), but their focus was on whether norm enforcers favoured their own group or not (rather than whether in-group favouritism is considered as a social norm). It was

[5] We focus on social norms rather than formal legal and judicial institutions because it is very difficult to observe exactly when in-group favouritism takes place. Therefore, formal sanctions through the judicial systems, which require verifiable evidence, are usually not effective at sanctioning this behaviour.

found that when the norm violator was from the same group as the enforcer (whilst the victim was an out-group) he received less punishment compared to an out-group norm violator. Furthermore, in these studies there was an *a priori* assumption on what kind of norm should be applied in a given situation. For example, in a dictator game, it is clear that a predominant norm is egalitarian sharing norm (equal split) between oneself and another person (Bernhard et al., 2006a; b). Similarly, in a Prisoner's Dilemma and a public goods game, the cooperation norm is usually assumed to be the predominant norm (Fehr and Gachter, 2000; Goette et al., 2006). However, it is more difficult to gauge what kind of norm should be enforced when there are multiple groups.

We design a new two-stage allocation experiment to address this question. In the first stage (decision stage), a decision-maker (the DM henceforth) decides how to allocate a fixed amount of money to the in-group and the out-group members (minimal groups). In the second stage (punishment stage), other players can observe the DM's choice and decide whether to punish the DM by reducing her payoff at a cost. The punishers can either be the in-group; the out-group; or a third-party who belongs to no group but observes the DM's choice. Only one type of punisher can punish the DM at a time. It is important to note that the in-group and the out-group punishers have been referred to as 'second-party punishers' (Fehr and Fischbacher, 2004) because they are directly affected by the norm violation. The third-party punishers, on the other hand, are not directly affected by the norm violation. However, since not everyone in the society is usually affected by a violation of a particular norm, if the second-party punishers impose sanctions, a very limited number of social norms would be enforced (Fehr and Fischbacher, 2004). Hence, there is a need for third-parties to enforce social norms in order for them to be sustained in the long-run (Bendor and Swistak, 2001).

Using this experimental design, we address the following specific research questions: *Do people who favour their own group get punished and if so, by whom? Is there a consensus amongst the punishers on what kind of norm should be enforced? Does the threat of punishment affect the DM's decision whether to favour her group?* To the best of our knowledge, our study is the first to systematically examine what kind of social norm governs in-group favouritism behaviour. We find that the punishers' choice of norm depends on their *own group identity*. If they belong to the in-group, then in-group favouritism is considered as acceptable. If they belong to the out-group, they are willing to punish in-group favouritism behaviour. Finally, third-party punishers do not seem care about in-group favouritism and thus, are not willing to punish. Our results shed light on the effectiveness of norm enforcement by showing that, unlike selfish or opportunistic behaviours, in-group favouritism is not always considered a violation of a social norm. Moreover, when group identities are taken into account, it can be ambiguous what the

predominant norm is, since punishers from different groups do not always agree on what type of social norm should be enforced. Our results also help to explain why in-group favouritism is widespread across societies.

The rest of the paper is organised as follow. Section 2 explains our methodological contributions to the literature. Section 3 describes the experimental design. Section 4 provides a theoretical framework that motivates our research hypotheses. Section 5 reports the behavioural results. Section 6 examines econometrically whether individual characteristics and attitudes influence the decision whether to favour one's own group in our experiment. Section 7 concludes.

2 Methodological contributions

Our experiment makes methodological contributions, which are worth noting separately. First, our design differs from the vast literature on in-group favouritism in social psychology and experimental economics in that it minimises the scope for ambiguous interpretation of observed in-group favouritism behaviour. It has been argued that in-group favouritism observed in the previous minimal group experiments did not reveal the true parochial or in-group bias tendency because the DMs bore no cost, regardless of their decisions. Instead, what appeared to be a preference for in-group favouritism in previous minimal group studies was likely to be based on the expectation of in-group reciprocity (Yamagishi and Kiyonari, 2000; Bernhard et al., 2006a; b). To address this problem, we use a one-shot anonymous game in which each subject knows only his or her own group identity and thus, reciprocation is not possible. In addition, we also test both a situation where the DM bears no cost (the baseline) and a situation where there is a potential cost to the DM if she deviates from the norm (the punishment treatments). By testing both conditions in the same design, we are also able to examine whether the in-group favouring decision is influenced by the expected cost (of punishment) or not.

Closely related to our paper are the studies by Bernhard et al., (2006a; b) which examined the parochial nature of altruistic punishment by third-party punishers. However, our design differs from theirs in a number of important respects. Firstly, in their design the DM divided the money between *herself* and another subject who either belonged to the same group or to a different group. This is contrary to our game where the allocation decision only affects the payoffs of *the other* in-group members and the out-group members. The DM's own payoff is not affected by her decision since she is not allowed to take any share of the in-group payoff. Secondly, in a standard dictator game used in their experiment it is clear that an equal split (50-50) is considered as the fairness norm, whereas the notion of fairness can be different when the

allocation decision is concerned with *other* people who come from different groups as in our design. Finally, their objective was to examine whether the third-party punishers acted more favourably towards the in-group when (i) the norm violators came from the in-group; and (ii) when the victims of the norm violation came from the in-group. This is not what we do here. Our objective is to examine whether the DMs who favour their own group get punished or not and by whom; and whether there is agreement amongst the punishers on the type of norm that they want to enforce.

Another related study is the paper by Goette et al., (2006) which investigated whether randomly assigned groups helped foster non-selfish cooperation and punishment of norm violation within the group. The objectives of their study are very similar to those of Berhard et al., (2006a; b). The main difference is that they used randomly assigned groups rather than naturally occurring groups and thus, the authors argue that the behaviours observed could be attributed to group membership rather than to other factors such as demographics or culture. Using a simultaneous prisoner's dilemma game, they found that subjects were significantly more cooperative when they interacted with a member of their own group; and the third-party punishers were more willing to enforce a norm of cooperative behaviour (which was assumed to be the predominant norm in their study) when the victim of defection was from the third-party's group. In addition to the fact that our objective is very different from theirs, the results of our paper also arise in a very different context. In the Prisoner's dilemma, a self-regarding player has a strict incentive to defect, so the threat of punishment must be sufficiently strong to induce a player to cooperate. In our design, any *threat* of punishment is enough to induce a self-regarding DM to abide by a social norm. Our theoretical model and experimental design are more sensitive to small amounts of other regarding preferences.

3 Experimental Design

In this section, we explain our experimental design in more detail. At the beginning of the experiment, seven subjects were randomly matched and divided into two groups: four subjects were randomly assigned to group A (the in-group) and three subjects to group B (the out-group). The reason for assigning four subjects to group A was because one person from this group would be randomly selected to be the DM. We wanted to make sure that there were equal numbers of the 'recipients' of the allocation decision in each group i.e. group A consisted of one DM and three in-group recipients and group B consisted of three out-group recipients. This is because in our design, the amount allocated was for *each* member of the group and not to the group as a whole (in order to ensure that the payoff was distributed equally within each group),

hence the number of people in each group could affect the DM's decision. For example, if group sizes were not the same, choosing option *d* would no longer yield equality in the total payoff. Group size could also have an independent effect on the DM's decision whether to favour the in-group[6]. For example, if the in-group was in the minority (majority) relative to the out-group, the DM might be more (less) likely to favour her group. In this case, in-group favouritism would be driven by the differences in group sizes rather than group categorisation. By keeping the group sizes symmetric, we were able to ensure that the DM's decision was only influenced by group categorisation.

We used an experimental monetary unit called 'token' to represent the payoffs in all treatments, which was then converted into real currency (the British Pound) immediately at the end of the experiment[7] and the game was played only *once*. Each subject was given an initial endowment of 3,000 tokens[8], regardless of their group assignment, and was clearly instructed that they were not asked to allocate this initial endowment. The experiment consisted of two stages: the decision stage and the punishment stage.

3.1 The Decision Stage

In the decision stage, each member of group A was asked to individually make an allocation decision, whilst members of group B were just passive players. We only allowed one group to make the allocation decisions in order to control for reciprocity. If both groups could make the allocation decisions, in-group favouritism may be driven by the belief that the other group would favour their group rather than group identity *per se*.

Once all members of group A made their decisions, the experimenter rolled a die to randomly select one person to be the decision-maker (DM).[9] The DM was not allowed to take any share from her group's payoff, but instead received a fixed payoff from the experimenter[10].

---

[6] We explore the effect of group size on in-group favouritism in a separate paper (in preparation).

[7] The exchange rate was 100 tokens = 7 pence. Subjects were clearly informed at the beginning of the experiment that this exchange rate was applied to their final payoff.

[8] This endowment was used as a buffer against bankruptcy because we allowed subjects to allocate a negative amount.

[9] It has been argued that this randomisation method used to select the DM *ex post* could induce 'generalised reciprocity' among the in-group members (Yamagishi and Kiyonari, 2000). To test whether this was the case in our design, we carried out a separate treatment, which had the same features as our original game, except that group A only consisted of *two* members. This increased the probability of being chosen to be the DM to 0.5 instead of 0.25 as in the original design and thus, if in-group favouritism was driven by generalised reciprocity, we should observe a higher proportion of in-group favouritism in the new treatment. However, we did *not* observe any increase in in-group favouritism. In fact, we observed a *decrease* in in-group favouritism when there were only two members in the group, which indicated that generalised reciprocity did not have a significant effect in our design.

[10] This design mimics a situation where a fixed salary is given to a public official by the government and the public official has to decide how to allocate public funds to different interest groups.

Therefore, the DM's decision did not directly affect her own payoff[11]. We set the DM's fixed payoff equal to the maximum amount that she could allocate to the other players (from a fixed choice set which we will discuss below), such that her decision would not be influenced by disadvantageous inequity aversion or envy (Fehr and Schmidt, 1999). An inequity-averse DM may not favour the in-group if it gives them a higher payoff than his own. While the members of group A were making allocation decisions, we elicited beliefs (non-incentivised) from the members of group B by asking them what they thought the DM would choose on average. This provides information about the expectation of in-group favouritism behaviour by the out-group.

The Choice Set

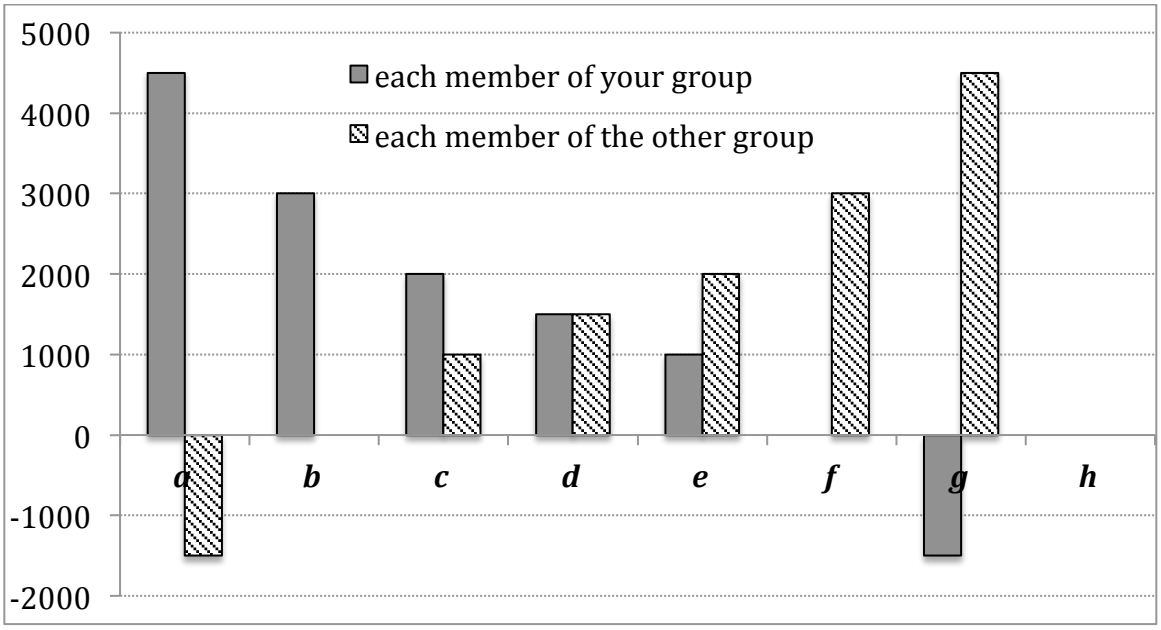
The allocation task was based on a fixed and symmetrical choice set. We used this fixed choice set to control for the subject's decision space and the stake size across all treatments. The choice set contained eight options – each differed in the amount allocated to each in-group and each out-group member. The first three options (*a, b, c*) represented different magnitudes of in-group favouritism as well as the cost imposed on each of the out-group member (*a* – the highest magnitude of in-group favouritism/cost {4,500; -1,500}[12]; *b* – medium magnitude {3,000; 0}; *c* – low magnitude {2,000; 1,000})[13]. Option *d* allocated equal amounts to both groups {1,500; 1,500} and in order to keep the choice set symmetrical, options *e* {1,000; 2,000}, *f* {0; 3,000}, and *g* {-1,500; 4,500} represented *out-group favouritism*. Finally, option *h* allocated zero token to both groups. Both options *h* and *d* allocated the same amount to each member of the two groups, but option *h* created a loss in social efficiency of 3,000 tokens. We used option *h* to disentangle fairness from efficiency concerns. The payoff distribution across the eight options is shown in Figure 1.

[11] But in the punishment treatments, self-interest may play a role in the form of a 'cost', which could be imposed on the DM by the punishers.

[12] The payoffs are presented here as {*in-group amount, out-group amount*}. The highest amount that each subject could receive from the allocation was 4,500 tokens, which was approximately 3.15 Pounds and the lowest amount was -1,500 tokens, which was around -1.05 Pounds, although none of the subjects left the experiment with a negative payoff because of the initial endowment.

[13] Lower-case letters were used in the choice set to prevent confusion with the group identities (A and B).

Figure 1: Payoff Distribution



Recall that our primary objective for assigning a fixed payoff to the DM was to control for self-interest, such that the DM's decision would not affected by disadvantageous inequity aversion (Fehr and Schmidt, 1999) or envy. But since the DM received a fixed payoff of 4,500 tokens, would option *a* appear more appealing than other options[14]? There are two things to consider here: inequity aversion and anchoring. If the DM was inequity averse and preferred equitable outcome for all players, choosing option *a* would not resolve the inequality problem because in our design the DM could not allocate 4,500 tokens to other players without giving some players extremely low payoffs (-1,500 tokens).

Could the fixed payoff of 4,500 tokens act as an 'anchor' on the DM's decision? We do not believe this was a problem. If anchoring was the dominant influence on the DM's decision then we should have observed that option *a* was chosen by the majority of the DMs in *all* treatments. But we did not observe this, particularly in the out-group punishment where only a small portion of the DMs chose option *a*. Furthermore, when we ran an additional treatment to test for generalised reciprocity (see *footnote 9*), the DM also received same fixed payoff (4,500 tokens), but the majority of the DMs did *not* choose option *a*, even though there was no punishment. Therefore, an anchoring effect was unlikely to be an important influence on the DMs' choices in our experiment.

[14] We thank an anonymous referee for pointing this out to us.

The Baseline Treatment (with no punishment)

The baseline treatment consisted of the decision stage only. After the DM was randomly selected, her choice (but not her personal identity) was revealed to all other players. The payoff for each subject was then calculated and shown on the computer screen and this concluded the experiment. The primary objective of the baseline treatment was to provide a benchmark against the punishment treatments in order to test whether there was any change in the DMs' behaviour when they were faced with the threat of punishment. In addition, the baseline treatment also allowed us to test whether we could replicate the previous experimental finding that people have a natural tendency to favour their own group.

However, our method differs from the standard minimal group paradigm used in many social psychology and experimental economics studies, which ask subjects to specify their preference for an object, usually a painting by the artists Paul Klee or Wassily Kandinsky (Tajfel et al., 1971; Chen and Li, 2009). They are subsequently categorised into either the 'Klee' or the 'Kandinsky' group. In our design, we used a much weaker notion of group identity. There was no shared preference and subjects did not know who were in their group or in the opposite group. They were only informed of their own group identity. Nevertheless, we maintained other criteria for the minimal group categorisation (Tajfel and Turner, 1986): (i) subjects are randomly assigned to non-overlapping groups; (ii) no social interaction takes place between subjects; (iii) group membership is anonymous; (iv) the decision task requires no link between a chooser's self interest and her choices. Moreover, the game was played only once and hence, reciprocity was ruled out as a motivating factor for in-group favouritism.

Furthermore, the baseline treatment also allowed us to examine whether in-group favouritism would be observed when the allocation decision affected the payoffs of *more than* one in-group member and one out-group member. It has been shown that as group size becomes larger, the amount of gift giving in a dictator game declines because increased group size reduces the value of the social surplus to the giver (Stahl and Haruvy, 2006; Andreoni, 2007). Therefore, we wanted to test whether having more than one other person in each of the recipient groups would change the DM's behaviour in our allocation game or not. If we observe in-group favouritism in our setting, it will provide further (and perhaps more unequivocal) support for the hypothesis that people have a natural tendency to favour their own group.

3.2. The Punishment Stage

In the punishment stage, the DM's choice was only revealed to the punishers, who then decided whether to punish the DM by deducting her payoff at a cost. Following the literature on social norms, punishment was costly to the punisher in order to ensure that norm enforcement was credible. In the *in-group punishment* treatment, the punishers were members of group A who were not selected to be the DM. At a cost of 100 tokens, each punisher could deduct 500 tokens from the DM's payoff[15]. The punishment decision was made individually and each punisher could choose one of three levels of punishment: deducting (i) 500 tokens (ii) 1,000 tokens or (iii) 1,500 tokens (at a cost of 100, 200, and 300 tokens respectively). The sum of the deductions was then subtracted from the DM's final payoff. Therefore, if all three punishers chose option (iii), the DM's payoff would be reduced to zero. In this treatment, the out-group members were just passive players[16]. Similarly, in the *out-group punishment* treatment, only the out-group members were assigned to be the punishers, whilst the in-group members were passive players and the same cost to punishment ratio was implemented. In both in-group punishment and out-group punishment treatments, there was no third-party punisher.

In the *third-party punishment* treatment, an independent third-party, who belonged to neither group A nor group B, was introduced to the game (as an eighth player). The DM's choice had absolutely no impact on the third-party's payoff. The third-party was given an endowment of 4,500 tokens from the experimenter, which was equivalent to the DM's fixed payoff in order to ensure that punishment decision was not influenced by disadvantageous inequity aversion (Fehr and Schmidt, 1999). In this treatment, only the third-party observed the DM's decision and decided whether to punish. Both the in-group and the out-group members were passive players. At a cost of 100 tokens, the third-party could deduct 300 tokens from the DM's payoff. The one-to-three cost to punishment ratio has been widely used in previous third-party punishment studies and has been shown to work effectively in enforcing a social norm (Fehr and Fischbacher, 2004b; Bernhard et al., 2006a;b). The punishment cost was slightly higher for the third-party compared to the in-group and out-group punishers because generally there is a higher opportunity cost for the third-party to 'get involved' when he is not directly affected by the norm violation (unlike the in-group or out-group members whose payoffs are directly affected). The higher cost also ensured that the third-party had to feel sufficiently strong about the norm that he wanted to enforce (Carpenter and Matthew, 2010). The third-party could spend up to 1,500

[15] Neutral frame was used in the instructions. We used the word 'deduct' instead of loaded words such as 'punish' or 'sanction'.
[16] We elicited beliefs of the out-group in the same way as in the baseline in all punishment treatments.

tokens (33% of their endowment) to reduce the DM's payoff to zero. Because there was only one third-party punisher in our design, he could impose a wider range of punishment on the DM (deducting between 300 and 4,500 tokens from the DM's payoff) than the in-group and out-group punishers who needed to coordinate their punishment effort if they wanted to reduce the DM's payoff to zero.

We only implemented *one* type of punishment in each treatment[17]. The reason for doing this was twofold. Firstly, we wanted to clearly distinguish the effect of the threat of punishment by *each type* of punisher on the DM's behaviour. If we allowed all three types of punishers to punish at the same time, we would not be able to identify which threat of punishment actually affected the DM's behaviour. Secondly, we wanted to test whether there was a consensus on a social norm that all three types of the punishers enforced without inducing a *free-riding problem* in the punishment stage (Casari and Luini, 2012). A free-riding problem could arise if we allowed all three types of punishers to punish at the same time because punishment was costly. If the punishers believed that others would enforce a norm, it would be in their interest to opt out of punishment in order to maximise their own payoff. If all types of punishers shared this belief, then no punishment would take place. By allowing only one type of punishment at a time, we were able to avoid the free-riding problem.

3.3 Experimental Procedures

The experiment was carried out in April 2009 in the UK at the University of Cambridge and University of Nottingham and was administered by z-Tree software (Fischbacher, 2007). We used the 'between-subject' design in order to maintain independence across treatments (Charness et al., 2012). After the subjects were seated, each was given a written instruction, which explained what they were to do in the experiment (see supplementary materials). They were

[17] Our design follows a large number of previous experimental studies, which examined second- and third-party punishments (Fehr and Fischbacher, 2004; Berhard et al., 2006 a; b; Goette et al., 2006; Casari and Luini, 2012; Leibbrandt and Lopez-Perez, 2012). These studies also allowed only one type of punisher to punish at a time (either the second-party or the third-party which is similar to our design). Furthermore, the fact that we observed variations in the punishment behaviours across all three treatments also suggests that *experimenter's demand effect* was not driving our results, particularly, in the third-party punishment treatment where only one person punished the DM. In previous studies, even though the third-parties usually punished less than the second-parties, more punishments were observed than in our study. Experimenter demand effects occur when subjects respond to a cue they receive from the experiment environment by behaving in a way, which they believe to be 'appropriate' for the task at hand (Zizzo, 2010). In our experiment, even if the subjects respond to the punishment 'cue' (since only one group can punish the DM at a time), it is still impossible for them to second-guess what the 'appropriate' punishment decision would be because that depends entirely on their own perception of the DM's behaviour. There is no 'right' or 'wrong' thing to do. Therefore, we do not think that our design and our subjects' behavours are influenced by experimenter demand effects. We thank an anonymous referee for raising this point.

given ten minutes to read the instruction and a verbal summary was also given afterwards. Therefore, the instruction was common information. They were also told explicitly how their payoffs were calculated. Each subject received a show-up fee of three British Pounds, which was added to the payoff at the end of the experiment. Subjects were paid privately at their seat and all seats were separated by partitions and thus, anonymity was retained. Each session lasted approximately 40 minutes and the average payoff was eight Pounds including the show-up fee.

4. Theoretical Framework and Research Hypotheses

This section describes a theoretical framework, which is adapted from Charness and Rabin (2002)'s model of social preferences, to motivate our research hypotheses and obtain predictions.

4.1. Model[18]

For a two-player setting, Charness and Rabin (2002) give a utility function whereby the utility of a player is a weighted sum of his own payoff (in monetary units) and the payoff of the other player. The weighting given to the other player's payoff may be negative or positive and is allowed to depend on whether the other player has a payoff which is higher or lower than that of the player in question. The weighting can also depend on whether the other player has 'misbehaved' in some manner, for example, by breaching a social norm.

Adapting the model of Charness and Rabin (2002)[19], we model each player's payoff as a weighted sum of the payoffs of all players whose payoffs are affected by his decisions, whether they are affected directly, or indirectly through the subsequent actions of others. For simplicity we model the three-player groups as individuals. That is, we have up to four types of player: a decision-maker (DM); a player in the same group as the DM (A); a player in the other group (B); a third-party punisher (3P). The payoffs of each are denoted π_{DM}, π_A, π_B, and π_3 respectively. Weightings given to the payoffs of others are ρ_{in}, σ_{in}, ρ_{out}, σ_{out}, where ρ, σ denote respective weightings for players poorer and wealthier than the player in question. 'In' and 'out' denote whether the weighting applies to a player in the same or in a different group to the player in question. The third-party is not in the same group as any other player.

The DM chooses an amount ξ_A to allocate to player A. $\xi_A \in [-1500, 4500]$. The amount allocated to player B is then given by $\xi_B = 3000 - \xi_A$. Following this decision, the punisher,

[18] We would like to thank Gary Charness for reviewing our model.
[19] For tractability we work with the model in the main body of the cited paper and not the more general model given in its appendix.

who may be A, B, or 3P, chooses an amount $\phi \in [0, 4500]$ that will be deducted from the DM's payoff. These two decisions determine payoffs. $\pi_{DM} = 4500 - \phi$. If A is the punisher, then $\pi_A = \xi_A - {}^{\phi}/_{5}$. If A is not the punisher, then $\pi_A = \xi_A$. Payoffs for B are similar. If 3P is the punisher, then $\pi_3 = 4500 - {}^{\phi}/_{3}$.

The DM's utility is given by:

$$u_{DM} = (1 - z_A - z_B - z_3)\pi_{DM} + z_A\pi_A + z_B\pi_B + z_3\pi_3$$

where $z_3 \equiv 0$ for the treatments without a third-party punisher, and

$$z_A = \begin{cases} \rho_{in} \text{ if } \pi_{DM} \geq \pi_A \\ \sigma_{in} \text{ if } \pi_{DM} < \pi_A \end{cases}, \qquad z_B = \begin{cases} \rho_{out} \text{ if } \pi_{DM} \geq \pi_B \\ \sigma_{out} \text{ if } \pi_{DM} < \pi_B \end{cases}, \qquad z_3 = \begin{cases} \rho_{out} \text{ if } \pi_{DM} \geq \pi_3 \\ \sigma_{out} \text{ if } \pi_{DM} < \pi_3 \end{cases}$$

Let the set of norms be a closed interval $\mathcal{N} \subseteq [-1500, 4500]$. These are the amounts that can be allocated to player A by the DM that are regarded as in accordance with typical behaviour. For example, a norm might be 'any split which gives the in-group player (A) at least as much as the out-group player (B)', that is $\mathcal{N} = [1500, 4500]$. Another possible norm would be 'an equal split' $\mathcal{N} = [1500, 1500] = \{1500\}$.

Define:

$$q = \begin{cases} 0 \text{ if } \xi_A \in \mathcal{N} \\ -1 \text{ if } \xi_A \notin \mathcal{N} \end{cases}.$$

That is, $q = 0$ when the DM acts in accordance with the norms, and $q = -1$ when the DM 'misbehaves'. In the latter case, the weighting of the DM's payoff will be reduced by an amount $\theta \geq 0$ in the utility function of the punisher.

The utilities of A and B when they are the punisher are:

$$u_A = (1 - w_{DM}^A - \theta q)\pi_A + (w_{DM}^A + \theta q)\pi_{DM}, \qquad w_{DM}^A = \begin{cases} \rho_{in} \text{ if } \pi_A \geq \pi_{DM} \\ \sigma_{in} \text{ if } \pi_A < \pi_{DM} \end{cases}$$

and

$$u_B = (1 - w_{DM}^B - \theta q)\pi_B + (w_{DM}^B + \theta q)\pi_{DM}, \qquad w_{DM}^B = \begin{cases} \rho_{out} \text{ if } \pi_B \geq \pi_{DM} \\ \sigma_{out} \text{ if } \pi_B < \pi_{DM} \end{cases}$$

As at any outcome, $\pi_3 \geq \pi_{DM}$, the utility of a third-party punisher is given by

$u_3 = (1 - \rho_{out} - \theta q)\pi_3 + (\rho_{out} + \theta q)\pi_{DM}$.

We assume that $\rho_{in} > \rho_{out}$, $\sigma_{in} > \sigma_{out}$, $\rho_{in} > \sigma_{in}$, $\rho_{out} > \sigma_{out}$. Further assume that $\rho_{in} > 0$ and $\frac{1}{6} > \rho_{in}, \sigma_{in}, \rho_{out}, \sigma_{out} > -\frac{1}{6}$. The last set of inequalities ensures that a player always places a weight of at least $\frac{1}{2}$ on his own payoff and that no DM will ever be willing to sacrifice his entire payoff to reduce the payoffs of out-group members.

4.2 Predictions

In the absence of punishment, the game consists of a single decision by the DM. It is always the case that $\pi_{DM} \geq \pi_A$, $\pi_{DM} \geq \pi_B$, therefore $z_A = \rho_{in}$, $z_B = \rho_{out}$. Then, because $\rho_{in} > \rho_{out}$, the optimal choice for the DM is to choose $\xi_A = 4500$, i.e. allocating everything to the in-group member[20]. As there is no punishment, it must then be that $\pi_A = 4500$, $\pi_B = -1500$, $\pi_{DM} = 4500$.

Hypothesis 1. *If minimal group categorisation leads to a positive charitable concern for the in-group* ($\rho_{in} > 0$) *which is larger than charitable concern for the out-group* ($\rho_{in} > \rho_{out}$)*, in-group favouritism will be observed: the DM will allocate a higher amount to the in-group than to the out-group.*

The game with punishment is a two-period game and is solved for subgame perfect equilibrium. Detailed derivations of all equilibria described in this section can be found in Appendix A. First, we consider the third-party punishment treatment. The third-party can sacrifice 1 unit of payoff in order to reduce the payoff of the DM by 3 units, subject to $\pi_{DM} \geq 0$. This leads to the condition that the third-party wishes to punish as much as possible if:

$\rho_{out} + \theta q < -\frac{1}{2}$

[20] Experimenter's demand effect is not an issue in our design since there is no compelling reason why the subjects should expect the experimenter to want them to favour their own group, especially when punishments are introduced (Zizzo, 2010).

and will not punish at all if the inequality is reversed. Note that our assumptions on charitable concern guarantee that $\rho_{out} > -\frac{1}{2}$, therefore a norm violation must occur ($q = -1$) for the third-party to want to punish. If $\rho_{out} - \theta < -\frac{1}{2}$, then in equilibrium the third-party will punish to the maximum extent possible if and only if there is a norm violation. That is, $\phi = 4500$ if $\xi_A \notin \mathcal{N}$, and $\phi = 0$ if $\xi_A \in \mathcal{N}$. Consequently, the DM will not violate the norm. Conditional on not violating the norm, the DM will give as much payoff as possible to the in-group player. That is, $\xi_A = \max\{\xi : \xi \in \mathcal{N}\}$. If $\rho_{out} - \theta > -\frac{1}{2}$, then the third-party will never punish, even if there is a norm violation, so that DM will give as much as possible to the in-group player. That is, $\phi = 0$ and $\xi_A = 4500$.

Hypothesis 2. *In the third-party punishment treatment, under the conditions of Hypothesis 1, then in-group favouritism will be observed unless (i) the norm does not allow in-group favouritism,* $\mathcal{N} \cap (1500, 4500) = \emptyset$, *and (ii) the third-party cares enough about the norm violation (*θ *is large enough* that $\rho_{out} - \theta < -\frac{1}{2}$).

Now consider the case where the punisher is the in-group member (player A). Player A can sacrifice 1 unit of payoff in return for reducing the payoff of the DM by 5 units, subject to $\pi_{DM} \geq 0$. This leads to the condition that A wishes to punish if:

$w_{DM}^{A} + \theta q < -\frac{1}{4}$

Note that as $w_{DM}^{A} \in \{\rho_{in}, \sigma_{in}\}$ and $\rho_{in} > \sigma_{in}$, player A will not necessarily punish as much as possible when he does punish. It is possible that punishment will stop at the point at which π_A becomes equal to π_{DM}. There are three cases to consider. If $\sigma_{in} - \theta > -\frac{1}{4}$ then A will never punish ($\phi = 0$) so the DM will give as much as possible to A ($\xi_A = 4500$). If $\sigma_{in} - \theta < -\frac{1}{4}$, then A will punish as much as possible ($\phi = 4500$) when a norm violation occurs, and will not punish otherwise. Consequently, the DM will give A as much as possible, conditional on not violating the norm ($\xi_A = \max\{\xi : \xi \in \mathcal{N}\}$).

The third case is the most interesting. If $\sigma_{in} - \theta < -\frac{1}{4}$ and $\rho_{in} - \theta > -\frac{1}{4}$, then when a norm violation occurs, A will punish until his payoff is equal to the payoff of the DM, or as much as possible if he cannot punish that much. That is, if $\xi_A \notin \mathcal{N}$, then $\phi = \min\{\frac{5}{4}(4500 -$

$\xi_A), 4500\}$. If $\xi_A \in \mathcal{N}$, then $\phi = 0$. Consequently, the DM will choose $\xi_A = 4500$ whether or not this is a norm violation. Player A will not punish as he already obtains the same payoff as the DM. Norm violations can occur in equilibrium.[21]

Hypothesis 3. *Under the conditions of Hypothesis 1, in-group favouritism will be observed in the in-group punishment treatment.*

Hypotheses 4 and 5 pertain to out of equilibrium play. Such play could arise from misunderstanding about the prevailing norm. That is, the DM may think he is choosing an action in $\mathcal{N}$, which is not in fact in $\mathcal{N}$ as understood by the punisher. From the above discussions of third-party and in-group punishment, we have the following prediction.

Hypothesis 4. *Any observed punishment will either be (i) severe, with the punisher punishing as much as possible, or (ii) egalitarian, with punishment equalizing as much as possible the payoffs of the punisher and the DM.*

When the punisher is the out-group player, results are similar to the case with an in-group punisher, with the difference that in the final case, the DM will never violate the norm, and will give A as much as possible conditional on this ($\xi_A = \max\{\xi: \xi \in \mathcal{N}\}$). Note that due to $\rho_{in} > \rho_{out}$, $\sigma_{in} > \sigma_{out}$, for a given value of θ, $\sigma_{in} - \theta < -\frac{1}{4}$ implies $\sigma_{out} - \theta < -\frac{1}{4}$, and $\rho_{in} - \theta < -\frac{1}{4}$ implies $\rho_{out} - \theta < -\frac{1}{4}$. Referring to the cases analysed above, this means that punishment for norm violations is always at least as harsh with an out-group punisher as with an in-group punisher.

Hypothesis 5. *Punishment for norm violations should be more frequent with out-group punishers than with in-group punishers.*

[21] This result does, of course, depend on the definition of norm used here. The literature on adaptive dynamics and conventions (see Young, 1993) considers norms to be equilibria that are anticipated precisely because they have been played in the past, and hence a norm that is routinely violated will not remain a norm for long.

5 Behavioural Results

Three hundred and fifty-nine subjects took part in the experiment (63 in the baseline; 84 in in-group punishment; 84 in out-group punishment; and 128 in third-party punishment). Most of the subjects were undergraduate students randomly selected from different faculties via a self-recruiting system, ORSEE (Greiner, 2004). The average age was 20 years old, 56% were males and only a small proportion of our subjects studied economics. This socioeconomic profile was consistent across all treatment groups as shown in *Table 1*.

The overall behavioural patterns across all treatments can be summarised as follows. Of the total 196 allocation decisions across all treatments[22], 76% chose in-group favouring options (*a, b* or *c*) and 21% divided the money equally between the two groups. Only small proportions of the subjects chose other options. The subjects were asked to rate their preferences (how much they preferred each option in the choice set) to check whether their preferences were consistent with their choice. The results confirmed that they chose the option that they most preferred. There were 88 punishers across all treatments (36 in the in-group punishment, 36 in the out-group punishment, and 16 in the third-party punishment). Only 18 (20%) decided to punish and most of the punishment was implemented in the out-group punishment treatment (14 punishers). 71% of the total punishment amounts was for in-group favouritism behaviour and the rest was for the equal distribution option. No other decisions were punished.

6.1 RESULT 1 (Baseline): The majority of the DMs (81%) favoured their own group.

Of the 36 members of group A, 81% chose either *a*, *b* or *c*. The most popular choice was option *a* (42%), which gave the maximum amount of 4,500 tokens to each of the in-group members at a cost of 1,500 tokens to each member of the out-group. The second most popular choice was option *c* (22%), which allocated slightly more money to the in-group (2,000 tokens) than the out-group (1,000 tokens), followed by the equal distribution option *d* (19%). Only one person chose to allocate nothing to either group (option *h*). The high proportions of options *a*, *b*, and *c*, particularly option *a*, support *Hypothesis 1* that the majority of the DMs had larger charitable concern for the in-group members. The fact that not all DMs chose option *a* indicated that preference for egalitarian outcome was present, particularly those who chose option *d*. Nevertheless, our results further strengthen previous findings that people have a natural tendency to favour their own group, showing that such a tendency persists when group categorisation is

[22] We use the allocation decisions of *all* group A members rather than only those who were randomly selected to be the DM to increase the number of observations in our analysis.

very weak; self-interest and reciprocity amongst the in-group members are controlled for; and each group consists of more than one member.

6.2 RESULT 2 (In-group punishment): the majority of the DMs favoured their own group and only DMs who did *not* choose option *a* were punished.

There was a slight increase in the overall proportion of in-group favouring options (*a*, *b*, or *c*) to 85% (from 81% in the baseline) and a small drop in the equal distribution option to 13% (from 19% in the baseline). Within the in-group favouring options, the proportion of the DMs who chose option *a* increased from 42% in the baseline to 50% and there was also an increase in the proportion of the DMs who chose option *b* from 17% to 21%. Our results support *Hypothesis 3* that in-group favouritism should be observed when the in-group has the power to punish.

For the punishment behaviours, only *three* out of the thirty-six in-group punishers (8%) decided to punish the DMs. The punishments were for options *b*, *c*, and *d*, whilst option *a* received no punishment. The fact that only small numbers of punishments were observed supports the equilibrium prediction of no punishment in equilibrium. That 50% of the DMs chose option *a* is consistent with the norm being one of in-group favouritism. Moreover, that punishment was only observed for choices other than option *a* suggests that punishment arising in out of equilibrium play could have arisen from misunderstanding about the prevailing norm. That is, the DM in such cases may have believed the norm to be other than *a*, with the punishers believing the norm to be *a*. With reference to the prediction of *Hypothesis 4*, observed punishment in one of the three cases was severe (option *c*), and observed punishment in the other two cases (options *b* and *d*) maybe driven by egalitarianism under the assumption that the punisher expected that the other in-group punishers would punish similarly to himself. For example, in the case that option *b* was punished, the punishment amount was 500 tokens. If all in-group punishers decided to deduct this amount from the DM's payoff (total punishment =1,500 tokens), the DM would be left with 3,000 tokens, which is equivalent to the amount that option *b* allocates to each of the in-group member.

Table 1: Allocation and Punishment decisions

Variables	Total	Baseline	In-group Punishment	Out-group Punishment	Third-party Punishment
Total subjects	359	63	84	84	128
Average Age	19.7	19.4	19.6	19.8	19.8
Male	56%	51%	55%	65%	52%
Undergraduate	99%	98%	98%	100%	100%
Study Economics	14%	19%	12%	13%	13%
Total allocation decisions	196	36	48	48	64
Option a {4,500: -1,500}	39%	42%	50%	27%	38%
Option b {3,000; 0}	18%	17%	21%	13%	20%
Option c {2,000; 1,000}	19%	22%	15%	13%	27%
Total In-group Favouritism	76%	81%	85%	52%	84%
Option d {1,500; 1,500}	21%	19%	13%	38%	16%
Option e {1,000; 2,000}	1%	0%	0%	2%	0%
Option f {0; 3,000}	1%	0%	0%	4%	0%
Option g {-1,500; 4,500}	1%	0%	0%	4%	0%
Option h (0; 0}	1%	3%	0%	0%	0%
No. of punishers	88	-	36	36	16
No. of punishment decisions	18	-	3	14	1
Total punished amounts (tokens)	18,900	-	3,000	13,500	2,400
No. punish option a (amounts)	4 (5,000)	-	0	4 (5,000)	-
No. punish option b (amounts)	5(6,400)	-	1 (500)	3 (3,500)	1 (2,400)
No. punish option c (amounts)	3 (2,000)	-	1 (1,500)	2 (1,500)	-
No. punish option d (amounts)	6 (4,500)	-	1 (1,000)	5 (3,500)	-
No. punish option e (amounts)	-	-	-	-	-
No. punish option f (amounts)	-	-	-	-	-
No. punish option g (amounts)	-	-	-	-	-
No. punish option h (amounts)	-	-	-	-	-

To gain more insight into the punishers' expectation of the DM's choice, the punishers' own allocation choices in the decision stage were examined. It was found that most in-group punishers (31 out of 36) chose in-group favouring options themselves, including the three who did punish the DMs. This information was important as it indicated that most of the in-group punishers considered in-group favouritism as a social norm. It also helps to explain why only a small number of punishments were observed. Since 85% of the DMs already chose to favour the in-group and thus, there was no need to implement punishment.

6.3 <u>RESULT 3 (Out-group punishment)</u>: The threat of out-group punishment siginificantly *decreased* in-group favouritism behaviour; and the out-group punished the DMs who chose in-group favouring options (*a, b* and *c*).

We observed significant changes in the DMs' behaviours in this treatment compared to

the baseline. First, there was a significant drop in in-group favouring options: from 81% in the baseline to 52% (Two-sample Wilcoxon rank-sum test [H0: favouritism choices (baseline) = favouritism choices (out-group punishment)]: z = 2.7, Prob >|z| =0.007) with the largest drop in option *a* (15%). Secondly, there was a significant increase in the equal distribution option (*d*) from 19% in the baseline to 38% (Two-sample Wilcoxon rank-sum test [H0: choose *d* (baseline) = choose *d* (out-group punishment)]: z = -1.78, Prob >|z| =0.07). Thirdly, 10% of the DMs favoured the out-group by choosing options *e*, *f*, or *g*. Compared to the baseline; the overall change in the DMs' behaviours was statistically significant at 5% level (Two-sample Wilcoxon rank-sum test [H0: Choices (baseline) = Choices (out-group punishment)]: z = -2.4, Prob >|z| =0.02). Our results suggest that a significant proportion of the DMs anticipated that they would be punished if they *favoured* the in-group and thus, chose the equal distribution option (option *d*). Some DMs even went as far as favouring the out-group.

The fact that a significantly larger proportion of the DM chose option *d*, compared to the baseline, could be due to their anticipation that the out-group would enforce an egalitarian norm or at least a norm antipathetic to in-group favouritism. That is, from the perspective of the out-group members, it is possible that they expected to be treated at least as favourably as the other group now that they had the power to punish the DM. Since the DM weighs the payoffs of the in-group more positively than the payoffs of the out-group (as shown in the theoretical model and backed by the results from the baseline treatment), the optimal choice for the DM is to give as much as possible to the in-group without causing a norm violation. Therefore, an 'apparent' egalitarian outcome could arise.

There were thirty-six out-group punishers in this treatment, of which fourteen (39%) decided to punish. Nine punished DMs who chose to favour their own group (choosing either options *a*, *b*, or *c)*, five of whom imposed the most severe punishment level (deducting 1,500 tokens from the DM's payoff so that the DM's would be zero were all three out-group punishers to similarly punish). The average size of punishment when it occurred was highest when the DM had chosen option *a*, and decreased monotonically through the choice of *b*, *c* and *d* by the DM. Maximal punishment is consistent with *Hypothesis 4*. Lower levels of punishment are sometimes consistent with an egalitarian motive (although this could be due to negative reciprocity). However, this is not true for the lowest level of punishment when options *c* and *d* are chosen. The fact that we observed such low level punishments suggests a non-linearity in the preferences of the punisher, whereby he takes a diminishing satisfaction in punishing the transgressor of a norm. This would be consistent with behavior in the real world where initial punishment is often at a low level and contains a signaling motive.

Since the out-group were not allowed to make real allocation decisions (in order to control for reciprocity), we asked them to make a hypothetical allocation decision from the same choice set (non-incentivised), whilst the members of group A were making real allocation decisions. These data on the stated preferences of the out-group punishers give an indication of their anticipated norm. We found that of those in the out-group who chose to punish, 50% chose an option in the hypothetical task, which involved *strictly less* in-group favouritism than the eventual choice of the DM which they chose to punish. 36% chose an option with the same amount of favouritism and 14% chose an option with strictly more favouritism. Interestingly, these figures are reversed when it comes to the stated beliefs of the punishers about what the DM would do, with fully 57% expecting more favouritism than that which they eventually observed and chose to punish. That is, the punishers would seem to have ascribed egalitarian[23] intent to themselves that they did not expect to be similarly present in the DMs. However, even the ascription of egalitarian morality to the punishers does not explain the five people who punished the equal division option.[24]

Finally, our theoretical prediction was that punishment of norm violations should be more frequent with out-group punishers than in-group punishers (*Hypothesis* 5). In-group punishments were indeed significantly less frequent (only 3 punishments) than out-group punishments (14 punishments). Average magnitudes of punishment were similar (3,000 tokens over 3 punishments compared to 13,500 tokens over 14 punishments). Previous studies (Bernhard et al., 2006a; b; Goette et al., 2006) found that when the punishers shared a group identity with the norm violator (in-group punishment), they behaved leniently towards the norm violator by punishing less severely. However, in our experiment there was no general agreement across treatments about which social norm should be enforced, so the magnitude of in-group versus out-group punishment when a given norm is violated cannot be measured. In-group favouritism seemed to be considered a norm by the in-group punishers, whist it was sometimes punished by the out-group. This is why it is interesting to examine third-party punishers' behaviour.

[23] Another possible motivation for punishing in-group favouring behaviour could be negative reciprocity. Since in-group favouritism in our game imposed a direct cost on the out-group members and hence, it might be perceived as 'unkind' behaviour (Rabin, 1993) by the out-group.

[24] It is difficult to speculate what motivated the out-group to punish the equal distribution option. One possible explanation is that they might be seeking to maximise the gain from punishment (Xiao, 2013). Since the payoff from option *d* was less than those from options *e, f*, and *g*, the allocator was punished for not choosing the latter options, which would give higher payoffs for the out-group. An alternative explanation may be anti-social punishment i.e. they perceived the allocators who chose the equal distribution as 'do-gooders' (Herrmann et al., 2008).

6.4 RESULT 4 (Third-party punishment): There was no significant change in the DMs' behaviours and the third-parties did not punish the DMs.

In this treatment, we observed a slight increase in the proportion of the DMs who chose in-group favouring options to 84% (from 81% in the baseline) and a slight drop in the equal distribution option to 16% (from 19% in the baseline). Within in-group favouring options, the proportion of the DMs who chose option *a* was slightly lower (38%) compared to the baseline (42%), whilst the proportions of options *b* and *c* were slightly higher. However, these changes were not statistically significant. Since the third-parties do not stand to gain or lose from the DM's decision and the punishment is costly, they have to feel sufficiently strongly about the norm violation in order to punish (Fehr and Fischbacher, 2004, a; b; Bernhard et. al., 2006a; b; Goette et al, 2006; Carpenter and Matthews, 2007). In this treatment, only one person decided to punish the DM. Since the rest of third-party punishers did not punish, either the DMs behavior was consistent with the expected norm or the third party punisher did not care sufficiently about norm violation to punish. Our results support *Hypothesis 2* that in-group favouritism will be observed and the third-party will not punish.

In sum, across the differing treatments, there was no general agreement amongst the punishers on the social norm that should be enforced when in-group favouritism was observed. Whether in-group favouritism was considered as a violation of a social norm seemed to depend on the context of the interaction. That is, expected norms differ according to the identity of the punisher. Moreover, conditional on a given treatment, beliefs of what constitutes reasonable behavior differs across subjects. When the punishers belonged to the in-group, in-group favouritism seemed to be considered a norm. When the punishers belonged to the out-group, although considered acceptable by some punishers, in-group favouritism was considered a norm violation by other punishers and consequently punished. Punishments by the out-group were more frequent than punishments by the in-group. This is consistent with differing payoff weightings in utility functions for in-group and out-group payoffs. Furthermore, it is also consistent with there being higher variation in what is considered reasonable behavior in the out-group punisher treatment than in the in-group punisher treatment. Finally, the third-parties either considered in-group favouritism as a social norm or they did not care much about this behaviour. In the post-experimental questionnaire, we asked the third-party punishers to state reasons for their decisions and most of them said that they 'did not see the need to incur a cost to punish the DMs'. The fact that the third-parties were not willing to punish in-group favouritism helps to explain why this behaviour is widespread across societies. Our results also shed new light on the

effectiveness of altruistic norm enforcement by showing that when there is no clear consensus on which social norm should be enforced; altruistic norm enforcement does not work effectively.

6.5 Out-group beliefs

In all treatments, we elicited beliefs (non-incentivised) from the members of group B (the out-group members) by asking them to state what they thought the members of group A would choose on average. As shown in *Table 2*, the majority of the out-group members thought that they would favour their own group. In the baseline, 85% believed that group A would favour their own group and 48% thought option *a* would be the most popular option. Only 11% thought the DMs would divide the money equally between the two groups. In the in-group punishment treatment, almost all out-group members (97%) believed that the DMs would favour their own group with a 10% increase in the expectation that option *a* would be chosen. In the out-group punishment treatment, 89% believed the DMs would favour their own group, but less people thought option *a* would be chosen (36%). Finally, in the third-party punishment treatment the beliefs were similar to the baseline. Our beliefs data suggested that there was a general consensus among the out-group members that group A would favour their own group and would do so even when they were faced with the threat of punishment by the out-group.

Table 2: Out-group Beliefs

Variables	Total	Baseline	In-group Punishment	Out-group Punishment	Third-party Punishment
Total out-group beliefs	147	27	36	36	48
Option a {4,500: -1,500}	46%	48%	56%	36%	44%
Option b {3,000; 0}	21%	22%	22%	22%	19%
Option c {2,000; 1,000}	23%	15%	19%	31%	25%
Total In-group Favouritism	90%	85%	97%	89%	88%
Option d {1,500; 1,500}	8%	11%	3%	8%	10%
Option e {1,000; 2,000}	1%	0%	0%	0%	2%
Option f {0; 3,000}	-	-	-	-	-
Option g {-1,500; 4,500}	1%	0%	0%	3%	0%
Option h (0; 0}	1%	4%	-	-	-

7 Heterogeneity in In-group Favouritism Behaviour

In this section we take a closer look at the heterogeneity in the DMs' behaviours. In our experiment, not all DMs adjusted their behaviours when faced with the threat of punishment. For example, 52% of the DMs in the out-group punishment chose to favour their own group, despite knowing that the out-group could punish them. On the other hand, not all subjects favoured their own group either. In all treatments, there were always some subjects who chose the equal distribution option (*d*): 19% in the baseline, 13% in the in-group punishment treatment, 38% in the out-group treatment, and 16% in the third-party punishment treatment.

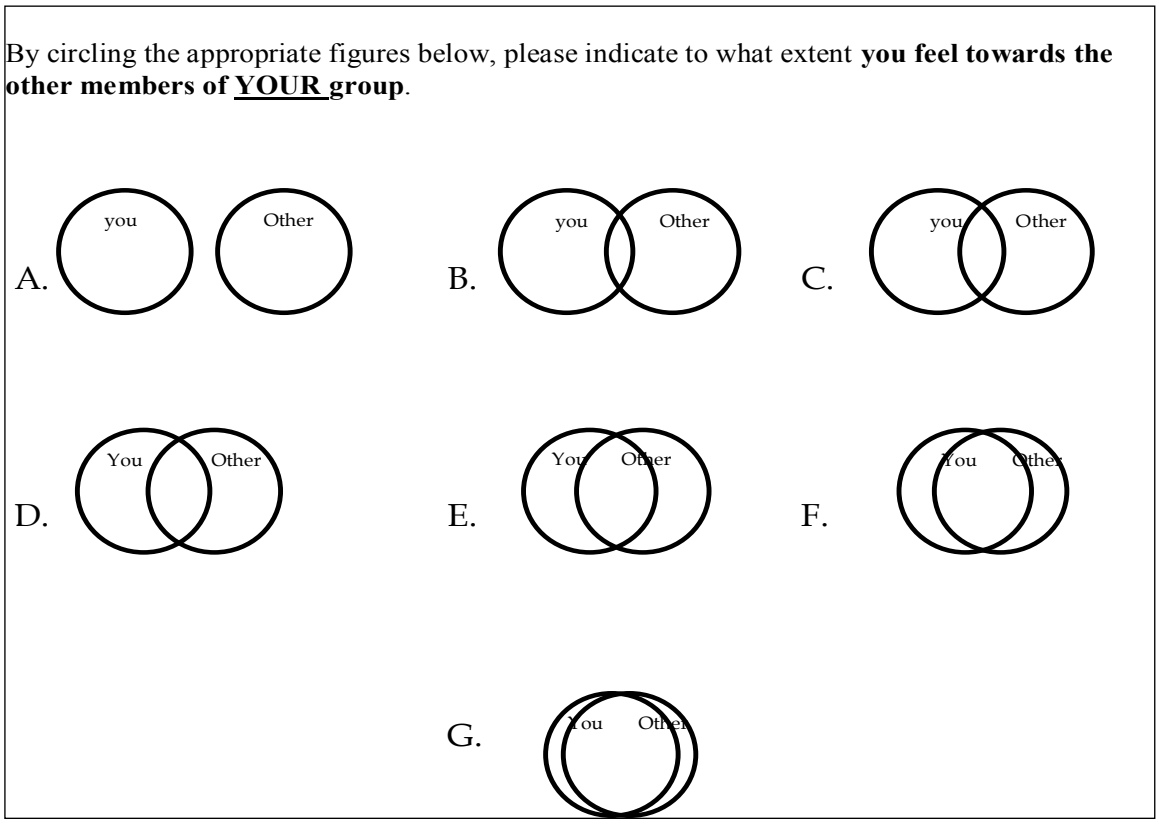
We empirically examine individual characteristics, which may help to explain this heterogeneity using a Probit model. The dependent variable is a binary variable that takes the value 1 if the DMs favoured their group (choosing options *a, b* or *c*) and is 0 otherwise. Our explanatory variables consist of (i) the treatment variables; and (ii) individual-specific characteristics and attitudes, including gender, age, saliency of group identity, generalised trust attitude, group equality attitude, and attitude towards bribery. We briefly explain these variables and how they are measured below.

7.1 Saliency of group identity

In order to check the internal validity of our randomised group assignment method, we used a psychometric test called '*the Inclusion of Other in the Self (IOS)*' scale (Aron et al., 1992; Cialdini et al, 1997), which measures the perceived self-other boundary overlap on a scale of 1 (very distant) to 7 (very closely overlapped), using a set of increasingly overlapping circles as shown in *Figure 2*. The subjects were asked to select only one pair of the circles, which they felt best described their relationship with the other in-group members or the out-group members.

We found that subjects who shared the same group identity (group A or B) felt closer towards their fellow group members compared to those in the other group. Across all treatments, the means IOS scale towards the in-group was 2.42 and that toward the out-group members was 1.95. We used the *difference* in the IOS scales (IOS in-group – IOS out-group) to measure the saliency of group categorisation: the larger the difference, the more salient the group identity. In the baseline, this difference was 0.46. In the in-group punishment treatment, it increased to 0.67. In the out-group and third-party punishment treatments, the differences were 0.42 and 0.40 respectively. These differences were all significant at 1% level (Wilcoxon signed-rank test), which confirmed that our randomised group categorisation method successfully created a distinction between the in-group and the out-group.

Figure 2: the Inclusion of Other in the Self (IOS) scale



7.2 Generalised Trust

Previous trust experiments have shown that people tend to be more trusting towards an in-group member (Glaeser et al., 2000; Fershtman et al., 2005; Falk and Zehnder, 2007). In our game, there may be an implicit trust amongst the in-group members to favour their own group and thus, the subjects who were more trusting in general might be more likely to favour the in-group. We used the widely cited question on generalised trust from the World Value Survey (WVS) which asked: "*Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?*" (1 = can't be too careful; 2 = most people can be trusted; 0 = I don't know). The mean trust attitude across all treatment was 1.24. The subjects in the baseline and the in-group punishment treatments also showed the same level of trust (1.24), whilst subjects in the out-group treatment were less trusting (1.08) and those in the third-party punishment treatment were slightly more trusting (1.35).

7.3 Group equality attitude

We used the 'Social Dominance Orientation (SDO) scale' (Pratto et al., 1994; Sidanius et

al., 2000) to measure individual's general attitudes towards group equality. A number of studies in experimental social psychology have shown that SDO scores predict political and economic conservatism, nationalism, anti-Black racism, and sexism (Jost, 2000). We selected eight statements from the SDO and our conjecture was that people who thought that group should be treated equally would be less likely to favour their own group and were more likely to choose equal distribution option. There were four statements which represent *positive group equality attitudes*[25]: (1) 'We should strive to make incomes as equal as possible' (55%); (2) 'Group equality should be our ideal' (69%); (3) 'We should do what we can to equalize conditions for different groups' (84%); (4) 'We would have fewer problems if groups were treated more equally' (74%). We also selected four statements which represented *negative group equality attitudes*: (1) 'It's OK if some groups have more of a chance in life than others' (23%); (2) 'To get ahead in life, it is sometimes necessary to step on other groups' (42%); (3) 'It's probably a good thing that certain groups are at the top and others are at the bottom' (33%); (4) 'Inferior groups should stay in their place' (5%). An index for '*positive group equality attitude*' was generated (with the highest number of factor loadings) using the Principal Factor Analysis (PFA) (Widaman, 1993) with varimax rotation method[26] and was used as one of the explanatory variables in the Probit model.

7.4 Attitude towards bribery

In-group favouritism, particularly nepotism, can be interpreted as a form of corruption. For example, if the DM is a public official who decides to allocate a government procurement contract, positions, or public resources to his friends rather the highest bidder or most efficient candidate, then in-group favouritism is a form of corruption. Therefore, we posit that people who agree that bribery is acceptable are more likely to engage in in-group favouritism and condone to this behaviour (do not punish). We used a question on political attitudes from the 2006 World Value Survey which asked whether the subjects agree or disagree (1= strongly disagree; 2=

[25] The percentage of the subjects who perceive each statement to be positive across all treatments is shown in parentheses.

[26] PFA is used when the research purpose is theory confirmation i.e. to determine if the number of factors and the loadings of measured variables (in this case, the SDO statements) on the factors conform to what is expected on the basis of pre-established theory (Kim and Mueller, 1978). The prior theory here is that there are two types of group equality attitudes: positive and negative. PFA analyses a correlation matrix in which the diagonal contains the communalities (equivalent to analysis of the covariance matrix) and thus, it accounts for the covariation among variables. The factors produced reflect the common variance of the variables. The main objective of the PFA is to find the least number of factors which can account for the common variance shared by a set of variables (Kim and Mueller, 1978). Analogous to Pearson's *r*, the squared factor loading is the percent of variance in that variable explained by the factor. Loadings above 0.6 are considered as 'high', whilst those below 0.4 are considered as 'low' (Hair et al., 1998; Raubenheimer, 2004).

Disagree; 3 =Neither agree nor disagree; 4= agree; 5= strongly agree; 0= I don't know) with the following statement: "*It is justifiable for someone to accept a bribe in the course of their duties*". The overall mean attitude towards bribery across all treatments was 1.58. It was the lowest in the baseline treatment (1.49) and the highest in the out-group punishment treatment (1.63). The mean attitude towards bribery in in-group punishment and third-party punishment were 1.55 and 1.59 respectively.

7.5 Empirical Results

Our main results, as shown in the first two columns of *Table 3* are as follows: (i) the threat of out-group punishment significantly reduced the propensity to favour one's own group by around 80%. The effect is robust when other controls were added to the model; (ii) in line with the behavioural results, the threats of in-group and third-party punishments did not have any effect on in-group favouring decision; (iii) age, gender, in-group saliency, generalised trust and the attitude towards bribery did not have any significant influence on in-group favoritism behaviour; (iv) Positive group equality attitude significantly reduced the propensity to favour own group by around 38%. The result confirmed our conjecture that people who had positive group equality attitude were less likely to favour their group and thus, helped explain the heterogeneity observed in the experiment.

We also ran Probit regressions with the punishment behaviours (1 = punish, 0 otherwise) as dependent variable and the same set of controls, as shown in the last two columns of *Table 3*. The main results are: (i) out-group punishers were significantly more likely to punish the DMs. The coefficient for this variable was large and was significant at 5% level; (ii) people who stated that bribery was not justifiable were about 42% more likely to punish, regardless of the treatment (recall that this variable was inversely scored: (1= strongly disagree; 5= strongly agree;); (iii) positive group equality attitude did not significantly influence punishment behaviour; however. We examined the pairwise correlation between positive group equality attitude and punishment behaviours and it was found to be very weak (6%). Our results, therefore, suggested that even though most of the subjects stated that group should be treated equally, they were not willing to incur a cost to punish in-group favouritism. It is worth noting that since the punishment observations are very small, we are cautious not to make an overarching conclusion from the punishment behaviour results.

8 Discussions and Concluding Remarks

In this paper, we examine whether in-group favouritism is considered as a norm in itself or as a violation of a different kind of norm, for example the egalitarian distribution norm. Using a new one-shot sequential allocation game where the decision-makers decided how to allocate a fixed sum of money between two minimal groups, controlling for self-interest, reciprocity, and disadvantageous inequity aversion, our main findings are as follows: (i) in line with previous studies, the majority of the decision-makers exhibited a tendency to favour their own group, even when the notion of a group was very weak; reciprocity and self-interest were ruled out; and each group consisted of three members; (ii) there was no general agreement amongst the norm enforcers on whether or not in-group favouritism was inconsistent with social norms and should be punished; (iii) Which social norm would be enforced appeared to be determined by the group identity of the punisher. When the punisher belonged to the in-group, in-group favouritism usually occurred and went unpunished. When the punishers belonged to the out-group, many of them chose to punish in-group favouritism. Finally, independent third-parties did not punish in-group favouritism. We find this result very interesting. It seems that there is something very different about in-group favouritism that sets it apart from selfish or opportunistic behaviours, which have been shown to be promptly punished by third-parties; (iv) we observed heterogeneity in the decision-makers' behaviours in the experiment and thus, examined econometrically whether individual characteristics and attitudes could explain this heterogeneity. We found that two variables played a significant role in influencing in-group favouring and punishment decisions. Positive group equality attitude significantly reduced the subjects' propensity to engage in in-group favouritism, whilst subjects who stated that accepting a bribe was not justifiable i.e. corruption was not acceptable, were more likely to punish the decision-makers who favoured their own group.

Our results provide an insight into the effectiveness of altruistic norm enforcement when group identity is taken into account. In our experiment, subjects stated generally that different groups should be treated equally in the self-reporting questionnaire, which could arguably be considered as a 'generally agreed' behavioural standard or a social norm. However, when they were put in a context in which the notion of 'groups' was made salient, even only by a very weak cue, their perception of in-group favouritism behaviour was influenced by their own group identity. Furthermore, our results also showed that when there was no general consensus on what kind of norm should be enforced, altruistic norm enforcement did not work. Future research may want to investigate further why the third-party punishers did not consider in-group favouritism as a violation of a social norm and since altruistic norm enforcement did not seem to work, other

incentive mechanisms, such as pecuniary and non-pecuniary rewards (given to the decision-makers who do not favour their own group) may need to be considered as an alternative mechanism to deter in-group favouritism.

Table 3:
Probit Models of Allocation and Punishment Decisions

	Dept. var = in-group fav		*Dept. var = punishment*	
Variables	**Model (1)**	**Model (2)**	**Model (1)**	**Model (2)**
Gender	0.29	0.40	-0.55	-0.72
	[0.22]	[0.24]	[0.35]	[0.39]
Age	0.12	0.03	0.03	0.49
	[0.04]	[0.05]	[0.12]	[0.12]
In-group Punishment	0.22	0.07	0.15	0.06
	[0.33]	[0.34]	[0.59]	[0.63]
Out-group Punishment	-0.77**	-0.80*	1.32**	1.09*
	[0.30]	[0.31]	[0.56]	[0.59]
Third-party punishment	0.16	0.10	omitted	omitted
	[0.31]	[0.32]		
In-group Saliency		0.07		-0.01
		[0.06]		[1.00]
Generalised trust		0.15		-0.26
		[0.17]		[0.27]
Positive group equality		-0.38**		0.27
		[0.16]		[0.25]
Attitude towards bribery		0.02		-0.42**
		[0.13]		[0.19]
cons	0.20	-0.41	-1.32	-1.85
	[0.99]	[1.06]	[2.34]	[2.53]
Obs	196	196	88	88
Pseudo R2	0.10	0.13	0.17	0.25

Reference:

Akerlof, G. and Kranton, R. (2000) 'The Economic of Identity'. *Quarterly Journal of Economics,* 115(3), pp. 715-753.

Akerlof, G. and Kranton, R. (2010) *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-being.* Princeton University Press.

Anderson, S. (2011) 'Castes as an Impediment to Trade'. *American Economic Journal: Applied Economics,* 3, pp. 239-263.

Andreoni J. (2007) 'Giving gifts to groups: How altruism depends on the number of recipients'. *Journal of Public Economics*, 91, pp. 1731–1749.

Aron A., Aron E. and Smollan D. (1992) 'Inclusion of Other in the Self Scale and the structure of interpersonal closeness'. *Journal of Personality and Social Psychology*, 63, pp. 596–612.

Bandiera O., Barankay I. and Rasul I. (2006) 'The Evolution of Cooperative Norms: Evidence from aNatural Field Experiment'. *Advances in Economic Analysis and Policy*, 6(2).

Bandiera, O., Barankay, I. and Rasul, I. (2009) 'Social connections and Incentive in the work place:Evidence from Personnel Data', *Econometrica*, 77(4), pp. 1047-1094.

Barr, A. and Oduro, A. (2002) 'Ethnic fractionalization in an African labour market'. *Journal of Development Economics*, 68, pp. 355-379.

Becker G. (1957) *The Economics of Discrimination*. The University of Chicago Press.

Belot, M. and van de Ven, J. (2011) 'Friendships and Favouritism on the Schoolground – A Framed Field Experiment'. *The Economic Journal*, 212, pp. 1288-1251.

Ben-Ner A., McCall B.P., Stephane M. and Wang H. (2009) 'Identity and in-group/out-group di↵erentiation in work and giving behaviors: Experimental evidence'. *Journal of Economic Behavior and Organization*, 72, pp. 153–170.

Bernhard H., Fehr E. and Fischbacher U. (2006a) 'Group Affiliation and Altruistic Norm Enforcement'. *American Economic Review*, 96(2), pp. 217–221.

Bernhard H., Fischbacher U. and Fehr E. (2006b) 'Parochial altruism in humans'. *Nature*, 442, pp. 912–915.

Billig M. and Tajfel H. (1973) 'Social categorization and similarity in intergroup behaviour'. *European Journal of Social Psychology*, 3, pp. 27–51.

Bolton G. and Ockenfels A. (2000) 'A Theory of Equity, Reciprocity, and Competition'. *American Economic Review*, 90(1), pp. 166–193.

Brandts J. and Charness G. (2011) 'The strategy versus the direct-response method: a first survey of experimental comparisons'. *Experimental Economics*, 14, p. 375398.

Brewer M. (1979) 'Intragroup bias in the minimal intergroup situation: a cognitive-motivational analysis'. *Psychological Bulletin*, 86, pp. 307–324.

Carpenter J. and Matthews P. (2009) 'What Norms Trigger Punishment'. *Experimental Economics*, 12(3), pp. 272–288.

Carpenter J. and Mattews P. (2010) 'Norm Enforcement: The Role of Third Parties', *Journal of Institutional and Theoretical Economics*, 166, pp. 239-258.

Casari M. and Luini L. (2012) ' Peer punishment in teams: expressive or instrumental choice?', *Experimental Economics*, 15(2), pp. 241-259.

Charness G., Rigotti L. and Rustichini A. (2007) 'Individual Behaviour and Group Membership'. *American Economic Review*, 97(4), pp. 1340–1352.

Charness G., Schram, A. (2013) 'Social and Moral Norms in Allocation Choices the Laboratory', Department of Economics, UCSB Working Papers qt0t39x0pt.

Charness G., Gneezy, U., and Kuhn, M. A. (2012) 'Experimental methods: Between-subject and within-subject design', *Journal of Economic Behavior & Organization*, 81(1), pp. 1-8.

Charness G. and Rabin, M. (2002) 'Understand Social Preferences with Simple Tests', *The Quarterly Journal of Economics*, 117(3), pp. 817-869.

Chen Y. and Li X. (2009) 'Group Identity and Social Preferences'. *American Economic Review*, 99(1), pp. 431–457.

Cialdini R., Brown S., Lewis B., Luce C. and Neuber S. (1997) 'Reinterpreting the Empathy – Altruism Relationship: When one Into One Equals Oneness'. *Journal of Personality and Social Psychology*, 73(3), pp. 481–494.

Elster J. (1989) *The Cement of Society: A Study of Social Order*. Cambridge University Press.

Fehr E. and Fischbacher U. (2004a) 'Social norms and human cooperation'. *TRENDS in Cognitive Sciences*, 8(4), pp. 184–189.

Fehr E. and Fischbacher U. (2004b) 'Third-party punishment and social norms.' *Evolution and Human Behaviour*, 25, pp. 63–87.

Fehr E., Fischbacher U. and Gaechter S. (2002) 'Strong reciprocity, human cooperation, and the enforcement of social norms'. *Human Nature*, 13.

Fehr E. and Gaechter S. (2000) 'Fairness and Retaliation: The Economics of Reciprocity'. *Journal of Economic Perspectives*, 14(3), pp. 159–181.

Fehr E. and Schmidt F. (1999) 'A Theory of Fairness, Competition, and Cooperation'. *Quarterly Journal of Economics*, 114(3), pp. 817–868.

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

Fisman, R. (2003) 'Ethnic Ties and the Provision of Credit: Relationship-Level Evidence from African Firms'. *Advances in Economic Analysis and Policy*, 3(1).

Gaube T. (2000) 'Group size and free riding when private and public goods are gross

substitutes'. Bonn Econonmic Discussion Papers 13/2000.

Global Competitiveness Report (2012), The World Bank.

Goeree J., Holt C. and Laury S. (2002) 'Private costs and public benefits: unraveling the effects of altruism and noisy behavior'. *Journal of Public Economics*, 83(2), pp. 255–276.

Goette L., Huffmann D. and Meier S. (2006) 'The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence using Random Assignment to Real Social Groups'. *American Economic Review*, 96(2), pp. 212-216.

Greiner K. (2004) 'An Online Recruitment System for Economic Experiments.' MPRA Paper, University Library of Munich, Germany.

Guth W., Levati M. and Ploner M. (2008) 'Social Identity and Trust: An Experimental Investigation'. *Journal of Socio-Economics*, 37(4), pp. 1293–1308.

Harris, D., Herrmann, B., Kontoleon, A. (2009). 'Two's Company, Three's a Group: The Impact of Group Identity and Group Size on In-group Favouritism', Centre for Decision Research & Experimental Economics, University of Nottingham, Working paper No. 2009-13.

Hoff K., Kshetramade M. and Fehr E. (2011) 'Caste and Punishment: The Legacy of Caste Culture in Norm Enforcement'. *The Economic Journal*, 121 (556), pp. F449-F475.

Isaac M., Walker J. and Williams A. (1994) 'Group size and the voluntary provision of public goods: experimental evidence utilizing large groups'. *Journal of Public Economics*, 54(1), pp. 1–36.

Naegelen F. and Mougeot M. (1998) 'Discriminatory public procurement and cost reduction incentives'. *Journal of Public Economics*, 67, pp. 349–367.

Predergast, C. and Topel, R.H. (1996) 'Favoritism in Organizations'. *Journal of Political Economy*, 104(5), 958-978.

Rabin M. (1993) 'Incorporating Fairness into Game Theory and Economics'. *American Economic Review*, 83, pp. 1281-1302.

Stahl D. and Haruvy E. (2006) 'Other-regarding preferences: Egalitarian warm glow, empathy, andgroup size'. *Journal of Behavior and Organization*, 61, pp. 20–41.

Schwab,K. (2012), Global Competitiveness Report. *The World Economic Forum.*

Tajfel H. and Turner J. (1986) 'The social identity theory of intergroup behavior'. In S. Worchel and L. Austin, eds., 'Psychology of Intergroup Relations', Chicago: Nelson-Hall.

Tajfel J., Billig M., Bundy R. and Flament C. (1971) 'Social Categorization in Intergroup Behaviour'. *European Journal of Social Psychology*, 1, pp. 149–178.

Xiao E. (2013) 'Profit-seeking punishment corrupts norm obenience'. *Game and Economic Behavior*, 77, pp. 321-344.

Yamagishi T. and Kiyonari T. (2000) 'The Group as the Container of Generalized Reciprocity'. *Social Psychology Quarterly*, 63(2), pp. 161–197.

Zantman W. (2002) 'Constitutional design and regional favouritism'. *Journal of Public Economics*, 4(1), pp. 71–93.

Zizzo D. (2010) 'Experimenter demand effects in economic experiments'. *Experimental Economics,* 13, pp. 75-98.

Appendix A. Theory – derivations

Appendix A.1. Third-party punisher (3P)

Substituting $\pi_{DM} = 4500 - \phi$ and $\pi_3 = 4500 - {}^{\phi}/_{3}$ into u_3, we obtain

$$u_3 = (1 - \rho_{out} - \theta q)\left(4500 - \tfrac{\phi}{3}\right) + (\rho_{out} + \theta q)(4500 - \phi) \quad \text{(A.1)}$$

and taking the derivative with respect to ϕ,

$$-\tfrac{1}{3}(1 - \rho_{out} - \theta q) - (\rho_{out} + \theta q) = -\tfrac{1}{3} - \tfrac{2}{3}(\rho_{out} + \theta q) \quad \text{(A.2)}$$

which is positive if $\rho_{out} + \theta q < -{}^{1}/_{2}$ and negative if $\rho_{out} + \theta q > -{}^{1}/_{2}$. Linearity implies that 3P will choose $\phi = 4500$ in the former case and $\phi = 0$ in the latter case. Hence, if $\rho_{out} - \theta > -{}^{1}/_{2}$ there will be no punishment even if a norm violation occurs. The DM's utility is then

$$\begin{aligned} u_{DM} &= (1 - z_A - z_B - z_3)\pi_{DM} + z_A\xi_A + z_B(3000 - \xi_A) + z_3\pi_3 \\ &= (1 - \rho_{in} - \rho_{out} - \rho_{out})4500 + \rho_{in}\xi_A + \rho_{out}(3000 - \xi_A) + \rho_{out}4500 \end{aligned} \quad \text{(A.3)}$$

and the derivative of this with respect to ξ_A equals $\rho_{in} - \rho_{out} > 0$, so it is optimal for the DM to choose $\xi_A = 4500$.

If $\rho_{out} - \theta < -{}^{1}/_{2}$ there will be punishment if a norm violation occurs, $\xi_A \notin \mathcal{N}$. The DM's utility when 3P sets $\phi = 4500$, making $\pi_{DM} = 0$, is bounded above in the following way

$$\begin{aligned} u_{DM} &= (1 - z_A - z_B - z_3)\pi_{DM} + z_A\pi_A + z_B\pi_B + z_3\pi_3 \\ &= z_A\xi_A + z_B(3000 - \xi_A) + z_3\left(4500 - \frac{4500}{3}\right) \\ &\leq z_A|\xi_A| + z_B|3000 - \xi_A| + z_3(3000) \\ &\leq \tfrac{1}{6}6000 + \tfrac{1}{6}3000 = 1500, \text{ as } |\xi_A| + |3000 - \xi_A| \leq 6000 \end{aligned} \quad \text{(A.4)}$$

The DM's utility when $\phi = 0$ is bounded below

$$\begin{aligned} u_{DM} &= (1 - z_A - z_B - z_3)\pi_{DM} + z_A\pi_A + z_B\pi_B + z_3\pi_3 \\ &= (1 - z_A - z_B - z_3)4500 + z_A\xi_A + z_B(3000 - \xi_A) + z_3 4500 \\ &= (1 - z_A - z_B)4500 + z_A\xi_A + z_B(3000 - \xi_A) \\ &> \tfrac{2}{3}4500 - \tfrac{1}{6}|\xi_A| - \tfrac{1}{6}|3000 - \xi_A| \qquad \left[\text{as } -\tfrac{1}{6} < z_A, z_B < \tfrac{1}{6}\right] \\ &\geq \tfrac{2}{3}4500 - \tfrac{1}{6}6000 = 2000 \end{aligned} \quad \text{(A.5)}$$

Therefore, the DM will never violate the norm and induce punishment in equilibrium. Given that his payoff is increasing in ξ_A as long as punishment is not induced, his optimal strategy is to choose $\xi_A = \max\{\xi : \xi \in \mathcal{N}\}$.

Appendix A.2. In-group Punisher

Substituting and taking the derivative as in (A.1) and (A.2) we obtain that A wishes to punish if

$$w_{DM}^A + \theta q < -\tfrac{1}{4} \tag{A.6}$$

and does not wish to punish if $w_{DM}^A + \theta q > -\frac{1}{4}$. Linearity implies that for any given value of w_{DM}^A there will either be no further punishment, or punishment will occur until the value of w_{DM}^A changes. This gives the three possibilities considered in the main text. If punishment never occurs or occurs up until $\pi_A = \pi_{DM}$, it follows immediately from u_{DM} that the DM's optimal choice is $\xi_A = 4500$, regardless of the norm. If, however, punishment is maximal ($\phi = 4500, \pi_{DM} = 0$) when norm violation occurs, then the DM's payoff when he violates a norm is bounded above

$$\begin{aligned} u_{DM} &= (1 - z_A - z_B)\pi_{DM} + z_A\pi_A + z_B\pi_B \\ &= z_A(\xi_A - \frac{4500}{5}) + z_B(3000 - \xi_A) \\ &\leq z_A|\xi_A - 900| + z_B|3000 - \xi_A| \\ &< \tfrac{1}{6}6900 + \tfrac{1}{6}3000 = 1650 \qquad \left[\text{as } z_A, z_B < \tfrac{1}{6}\right] \end{aligned} \tag{A.7}$$

The lower bound in (A.5) still holds, therefore the DM will not violate the norm and will choose $\xi_A = \max\{\xi : \xi \in \mathcal{N}\}$.

Appendix A.3. Out-group punisher

In a similar manner to in-group punishment, there are three cases. If $\sigma_{out} - \theta > -1/4$, then punishment never happens and the DM chooses $\xi_A = 4500$. If $\rho_{out} - \theta < -1/4$, then punishment happens to the maximum extent when the norm is violated, and as bounds (A.5) and (A.7) still hold, the DM will not violate the norm and will choose $\xi_A = \max\{\xi : \xi \in \mathcal{N}\}$. Finally, consider the case $\sigma_{out} - \theta < -1/4$, $\rho_{out} - \theta > -1/4$. In this case punishment is not greater than the amount required for $\pi_B = \pi_{DM}$. Specifically, when $\xi_A \notin \mathcal{N}$, the amount of punishment (up to a maximum 4500) solves

$$4500 - \phi = 3000 - \xi_A - \tfrac{\phi}{5}, \tag{A.8}$$

which gives

$$\phi = \min\{\tfrac{5}{4}(1500 + \xi_A), 4500\} \tag{A.9}$$

We already know from the previous cases that the DM will not violate a norm if this results in maximal ($\phi = 4500$) punishment. Consider a situation where the DM is choosing $\xi_A > \max\{\xi : \xi \in \mathcal{N}\}$ and is being punished at some $\phi < 4500$. We examine the benefits and costs to the DM of a marginal change in ξ_A. The DM's utility is

$$\begin{aligned} u_{DM} &= (1 - z_A - z_B)\pi_{DM} + z_A\pi_A + z_B\pi_B \\ &= (1 - z_A - z_B)(4500 - \phi) + z_A\xi_A + z_B(3000 - \xi_A - \frac{\phi}{5}) \\ &= (1 - z_A - z_B)(4500 - \tfrac{5}{4}(1500 + \xi_A)) + z_A\xi_A + z_B(3000 - \xi_A - \tfrac{1}{4}(1500 + \xi_A)), \end{aligned} \tag{A.10}$$

the derivative of which with respect to ξ_A is

$$\begin{aligned} &-(1 - z_A - z_B)\frac{5}{4} + z_A + z_B(-1 - \frac{1}{4}) \\ &< -\frac{2}{3}\frac{5}{4} + \frac{1}{6} + \frac{1}{6}\frac{5}{4} \qquad \left[\text{as} -\frac{1}{6} < z_A, z_B < \frac{1}{6}\right] \\ &= -\tfrac{11}{24} < 0 \end{aligned} \tag{A.11}$$

Therefore, the DM would wish to reduce ξ_A and to keep doing so until $\xi_A \in \mathcal{N}$.

Appendix B

Instructions

Welcome. You are now taking part in an economic experiment about decision-making financed by research foundations.

These instructions are solely for your private use. **It is prohibited to communicate with other participants during the experiment.** Should you have any questions, **please ask the administrator**.
If you violate this rule, you will be dismissed from the experiment and forfeit all payments.

During the experiment we will not speak in terms of Pound but in '**Token**'. During the experiment, your entire earnings will be calculated in 'Token'. At the end of the experiment the total amount of 'Tokens' which you have earned will be converted to Pounds at the following exchange rate:

100 Tokens = 7 pence

At the end of the experiment, your entire earning from the experiment plus the 3 Pounds on-time show-up fee will be paid to you in cash in private.

In the following pages, we describe the experiment in detail.

Detailed Information of the Experiment

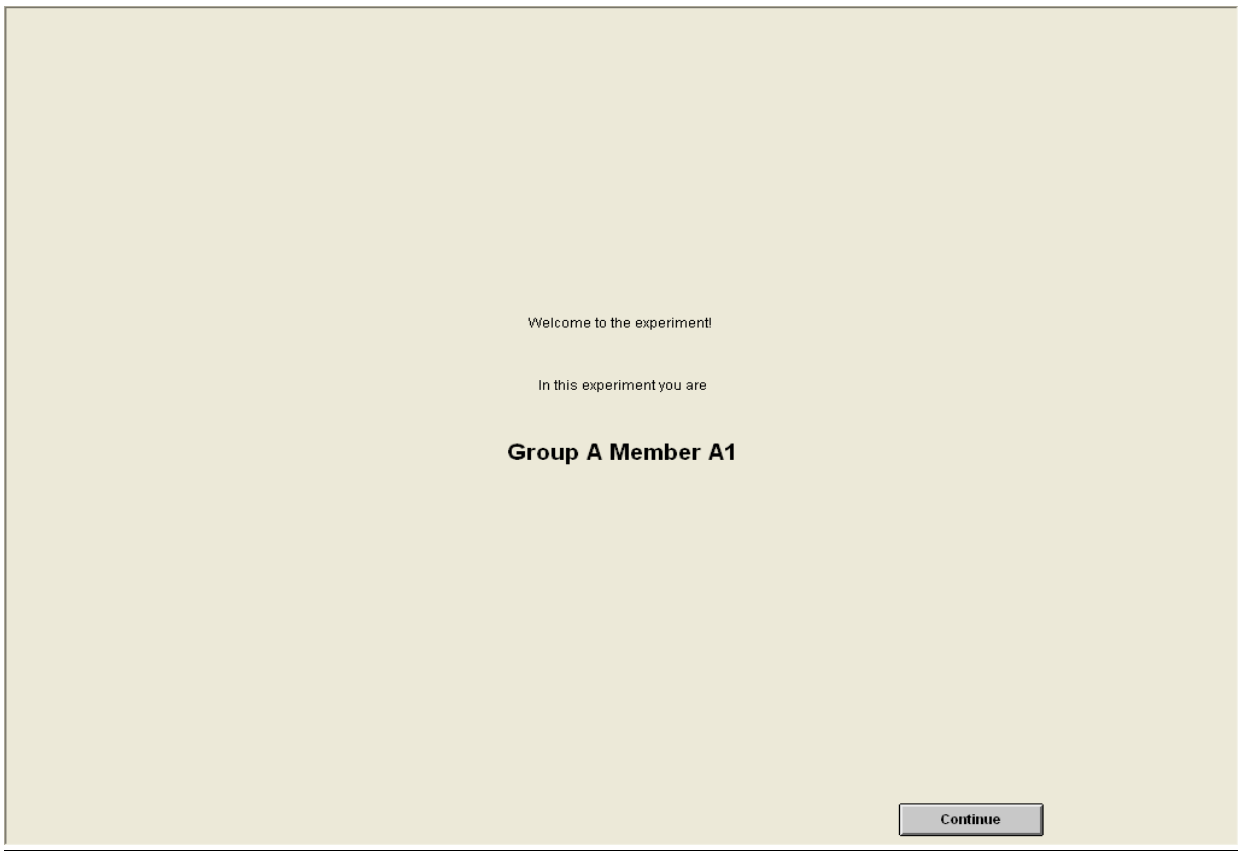
At the beginning of the experiment, all participants will each be given an initial lump sum of 3,000 Tokens. Furthermore, in this experiment there is an additional allocation decision, which can earn you and other participants extra payment. Please note that you are **not** allocating these initial 3,000 Tokens. You will be given a separate set of allocation options where each allocation option will either add to or subtract from the initial lump sum. **You will only make this allocation decision ONCE.**

In this experiment, participants are randomly divided into sets of seven participants whose identity you will never find out either before, during or after the experiment. In each set, **four** participants will be randomly grouped together to form a group called 'GROUP A' and **three** other participant will be grouped together to form 'GROUP B'. Within GROUP A, the group members will be assigned one of the following roles: A1, A2, A3, or A4.

Each member of **GROUP A** will make an allocation decision which can affect the incomes of **the other** GROUP A's members and the members of GROUP B. Each member of GROUP A will make this decision only **ONCE**.

The members of **GROUP B** will make **no** allocation decision, but will be asked to do different tasks which will be explained below. The role you are assigned will be shown on your screen at the beginning of the experiment.

For example if you are assigned to be in GROUP A and member no. A1, you will see the following screen.



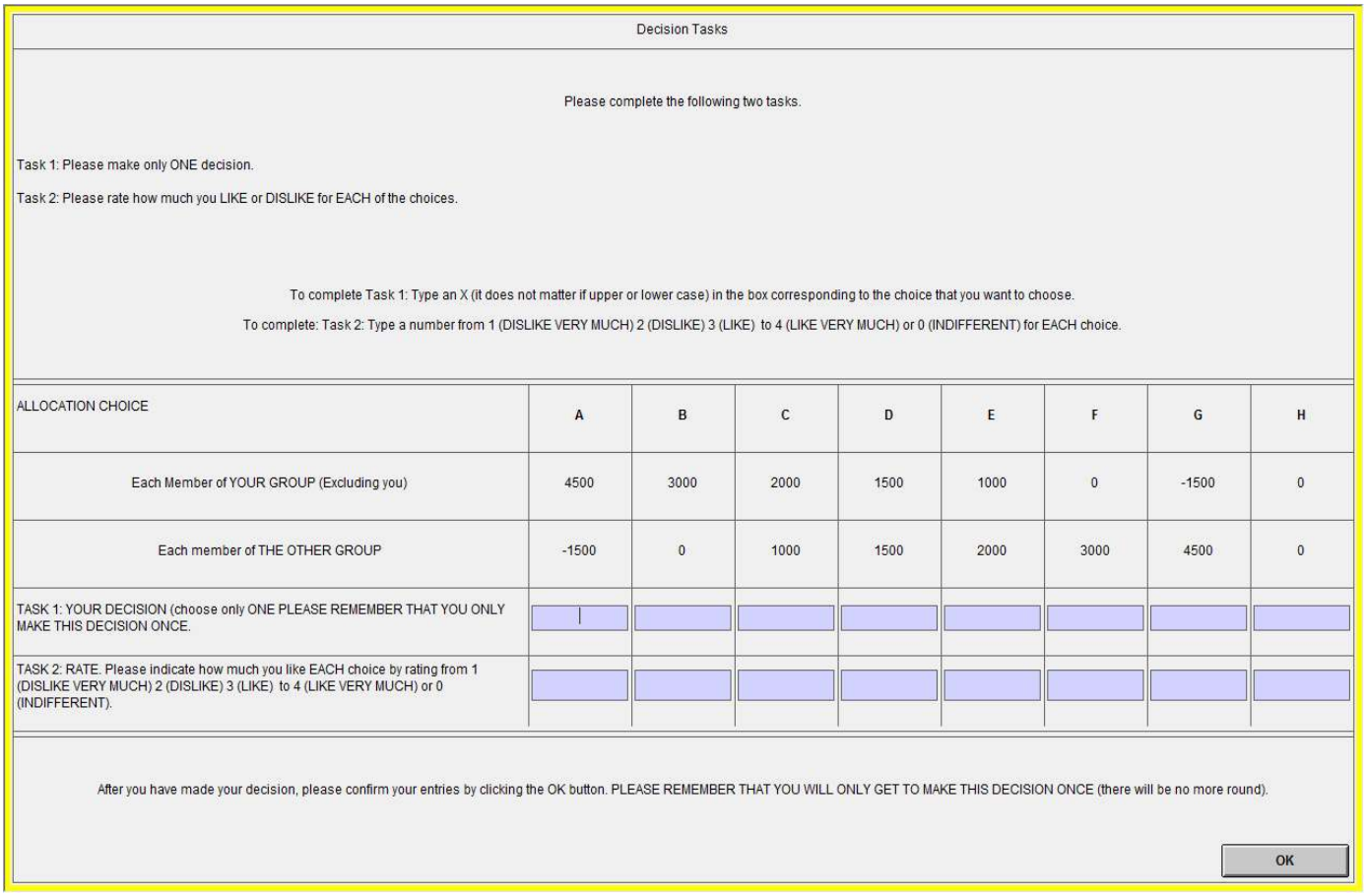
Members of Group A

If you are assigned to be in GROUP A, you will be asked to complete the following **TWO** tasks.

TASK 1: In task 1, on the screen you will see a set of eight different allocation options (an example of the allocation options is shown below). Each option allocates Tokens between your fellow GROUP A's members (**excluding you**) and the members of GROUP B. To indicate your decision, mark an 'x' (it does not matter whether it's lower or upper case) in the box under the option that you would like to choose. **Please choose only ONE option and please remember that you will only make this decision ONCE**.

TASK 2: In task 2, after you have made the decision, you are asked to **RATE EACH** option using the following scale: **1 (Dislike very much), 2 (Dislike), 3 (Like), 4 (Like very much), or 0 (Indifferent)**.

Example of Allocation Options



Decision Tasks

Please complete the following two tasks.

Task 1: Please make only ONE decision.

Task 2: Please rate how much you LIKE or DISLIKE for EACH of the choices.

To complete Task 1: Type an X (it does not matter if upper or lower case) in the box corresponding to the choice that you want to choose.

To complete: Task 2: Type a number from 1 (DISLIKE VERY MUCH) 2 (DISLIKE) 3 (LIKE) to 4 (LIKE VERY MUCH) or 0 (INDIFFERENT) for EACH choice.

ALLOCATION CHOICE	A	B	C	D	E	F	G	H
Each Member of YOUR GROUP (Excluding you)	4500	3000	2000	1500	1000	0	-1500	0
Each member of THE OTHER GROUP	-1500	0	1000	1500	2000	3000	4500	0
TASK 1: YOUR DECISION (choose only ONE PLEASE REMEMBER THAT YOU ONLY MAKE THIS DECISION ONCE.								
TASK 2: RATE. Please indicate how much you like EACH choice by rating from 1 (DISLIKE VERY MUCH) 2 (DISLIKE) 3 (LIKE) to 4 (LIKE VERY MUCH) or 0 (INDIFFERENT).								

After you have made your decision, please confirm your entries by clicking the OK button. PLEASE REMEMBER THAT YOU WILL ONLY GET TO MAKE THIS DECISION ONCE (there will be no more round).

OK

You and your fellow GROUP A members will each make the decision, but only **ONE** of the four decisions will be randomly selected. Once all the GROUP A members have completed both tasks and confirmed their decisions and ratings by clicking the OK button, the administrator will throw a dice which will determine whose decision will be selected:

If the dice shows number **1**, the decision of **GROUP A's member A1** will be selected.
If the dice shows number **2**, the decision of **GROUP A's member A2** will be selected.
If the dice shows number **3**, the decision of **GROUP A's member A3** will be selected.
If the dice shows number **4**, the decision of **GROUP A's member A4** will be selected.

If the dice shows numbers 5 or 6, no decision will be selected and the dice will be thrown again until it shows the numbers between 1 and 4.

The payoff of the GROUP A member whose decision **is** selected will **not** be affected by his/her decision. Instead, he/she will be given **a fixed payment of 4,500 Tokens** and his/her decision will be kept confidential. The payoff of the GROUP A members whose decision are **not** chosen, their payoff will be determined by the selected decision.

For example, if you are assigned a role of **A1 and** your decision **is** chosen (the dice shows number 1) your payoff **will not** be affected by your decision. However, if you are assigned a role of A1 but your decision **is not** chosen, your payoff **will** be affected by the decision made by the GROUP A member whose decision is chosen.

Once the decision is determined, you can see your payoff on the screen.

Members of Group B

If you are assigned to be in GROUP B, you will be asked to complete **TWO** tasks:

TASK 1: You will be asked to select an option which you think the decision-maker in GROUP A is most likely to choose. To indicate your decision, mark an 'x' (it does not matter whether it's lower or upper case) in the box under the option **that you think the decision-maker in GROUP A is most likely to choose.** Please choose only **ONE** option and you can only make this decision **ONCE.**

TASK 2: Suppose that you were in the position to decide how to allocate the Tokens between the members of your group (GROUP B) and the other group (GROUP A). **Which option would you choose?** To indicate your decision, mark an 'x' (it does not matter whether it's lower or upper case) in the box under the option you would like to choose. Please choose only **ONE** option. Examples of the decision screens for members of GROUP B are shown below.

Decision Tasks

Please complete the following task:

Task 1: Select the allocation that you think is most likely to be chosen by THE DECISION-MAKER IN GROUP A.

To complete Task 1: Type an X (does not matter if upper or lower case) in the box corresponding to the predicted choice.)

ALLOCATION CHOICE	A	B	C	D	E	F	G	H
GROUP A	4500	3000	2000	1500	1000	0	-1500	0
GROUP B	-1500	0	1000	1500	2000	3000	4500	0
TASK 1: A) Your PREDICTION (Please choose ONLY ONE).								

After you have made your decision, please confirm your entries by clicking the OK button. PLEASE REMEMBER THAT YOU WILL ONLY GET TO MAKE THIS DECISION ONCE (there will be no more round).

OK

Decision Tasks

Please complete the following task:

Suppose you were in the position to decide how to allocate Tokens between YOUR GROUP (GROUP B) and the OTHER GROUP (GROUP A). What would you do?

To complete Taks 2: Type an X (does not matter if upper or lower case) in the box corresponding to the choice that you want to choose.

ALLOCATION CHOICE	A	B	C	D	E	F	G	H
YOUR GROUP	4500	3000	2000	1500	1000	0	-1500	0
THE OTHER GROUP	-1500	0	1000	1500	2000	3000	4500	0
TASK 2: What would YOU do IF you were you were in the position to decide how to allocate Tokens between YOUR GROUP and the OTHER GROUP? (choose ONLY ONE).								

After you have made your decision, please confirm your entries by clicking the OK button.

OK

Once the dice is thrown and the decision is determined, the members of GROUP B will also see their payoffs on the screen.

After all players have completed their tasks, the administrator will ask you to complete a short questionnaire and will also distribute a document which you will need for the questionnaire.

After you have completed the questionnaire, the administrator will come over to your seat and give you the payment in private.

How the payment for each player is calculated is shown below:

Your payment

Members of GROUP A

If you are in GROUP A and your decision **is** chosen, your payment will be:
3,000 Tokens + 4,500 Tokens + £3 (on-time show-up fee)

If you are in GROUP A, but your decision is **not** chosen, your payment will be:
3,000 Tokens + the amount allocated to you by the decision-maker + £3 (on-time show-up fee)

Members of GROUP B

If you are in GROUP B, your payment will be:
3,000 Tokens + the amount allocated to you by the decision-maker in GROUP A + £3 (on-time show-up fee)

Now if you have any question, please raise your hand and the administrator will come and assist you.