# The Effect of Inequality Aversion on a Climate Coalition Formation: Theory and Experimental Evidence

Lin, Yu-Hsuan

Department of Economics, the Catholic University of Korea

January 2017

# Inequality-Averse Preference for International Environmental Agreements

Yu-Hsuan Lin

*Catholic University of Korea, Korea*

*Corresponding author at: 43, Jibong-ro, Wonmi-gu, Bucheon-si, Gyeonggi-do 420-743, Republic of Korea. E-mail address: yuhsuan.lin@catholic.ac.kr*

**Abstract**

This paper explores the individual incentives of participating in international environmental agreements (IEAs) with social preferences (also known as other-regarding preferences) in a static model through the experimental method. More specifically, the focus is on the impact of individual preferences of inequality-aversion. The experimental method has been used to capture actual decisions in a purified laboratory environment. Our theoretical prediction expects that players with high degree of inequality-averse preferences will violate the internal and the external constraints. As a consequence, the stable coalition formation may not necessary exists. The experimental outcomes show that a stable coalition is indeed very rare. Individual preferences on inequality-aversion do matter for coalition formation. However, highly inequality-averse subjects, are less likely to violate the internal constraint by leaving the coalition. Hence, the IEA formation is usually larger than the equilibrium formation.

*Key words:* Social preference, experimental design, international environmental agreement, inequality aversion
*JEL:* code: C91, H41, Q54

## 1. Introduction

International environmental agreements (IEAs) are typically viewed as coalitions of agents providing public goods (e.g., abatements of greenhouse gas emissions). Barrett (1994) provides a seminal study that positions 'self-enforcing' as a key incentive for providing public goods by participating and

interacting in IEAs. His key assumption of the absence of a supra-national body to structure an IEA leads him to suggest that participation is voluntary and all countries are free to enter or to withdraw from a coalition. While an IEA aims to maximise the aggregate net benefit, individual nonsignatories aim to maximise their own net benefit. In joining an IEA, signatories receive a reward from acceding to the agreement and avoid the punishment from withdrawing. Non-signatories may be penalised but also enjoy the free-riding benefit. The majority of the literature, however, follows D'Aspremont *et al.* (1983) who argue that a stable coalition has two constraints: the internal one where no signatory has any incentive to withdraw from the coalition; and the external one where no nonsignatory has any incentive to join the coalition.

In the existing literature, two issues still await to be addressed: the arguably unavoidable free-riding effect and a presumed egoistic preference.

Free-riding has largely been considered as the most important obstacle for the formation and existence of successful IEAs. This is the main reason why, a large IEA is not easy to be formed without any policing mechanism, in light of the Nash equilibria static game. However, recent experimental evidences on IEAs suggest that people are far less likely to free ride and more likely to cooperate than the theory suggests (Kosfeld *et al.*, 2009; Burger and Kolstad, 2009). But why this is so has not been well-explained by the models in the literature.

Furthermore, existing research findings on IEAs largely presume that an individual's preference is egoistic/selfish. In light of the Nash equilibrium, this implies that a rational agent would choose the highest payoff. The assumption has been widely employed in the majority of the theoretical studies of IEAs (e.g. Barrett, 1994, 2001; and Breton *et al.* 2010). However, recent experimental evidences have suggested that the assumption of egoistic preferences is not enough to explain individual decision makers' behaviours in an interactive game (Kosfeld *et al.*, *ibid*; Burger and Kolstad, *ibid*). These studies claim that people are far less likely to free ride and more likely to cooperate than the egoistic prediction assumes. Hence, social preference (or other-regarding preference) has been proposed in recent studies (e.g. Kolstad, 2014) to address this gap. The solutions to international environmental problems require cooperation and interaction between different nations at a global scale so as to prevent environmental or natural disasters or damages from happening. International cooperations are called for to deal with global issues. In such interactive game with common goal to minimise the loss of the society and environment, the assumption of a pure egoistic preference may

not be enough to capture players behaviours. This study follows this trend of thought and considers two types of other-regarding preference, namely inequality-aversion and altruism, to develop the model and experimental design.

Some have suggested to address this limitation by taking the role of other-regarding preferences (also known as social preferences) into account. Kosfeld *et al.* (*ibid*) employ the inequality-averse preference (proposed by Fehr and Schmidt, 1999) and confirm with laboratory-based evidence that when inequality-averse players exist, the coalition is no longer a Nash prediction, and the grand coalition becomes an expected equilibrium outcome. On the other hand, Kolstad (*ibid*) adopts Charness and Rabin's (2002) social preferences theory which suggest that agents mainly care about three things : private payoff, fairness in payoffs, and overall efficiency. In contrast to the finding of Kosfeld *et al.* (*ibid*), Kolstad argues that the size of an equilibrium of a coalition is smaller when social preferences exist.

Although the coalition formation with social preference has been examined in the literature, its influence on individual behaviours in an interactive coalition has not been fully explored. In other words, individual incentives for participating in a coalition are still unclear. This is partly due to the fact that economic models usually are based on several assumptions to reduce uncertainties and ambiguities. But these assumptions make capturing individual incentives difficult. For example, even with the assumption of heterogeneous agents, players were given the same payoff table in an experiment. There exist multiple equilibria and several possible coalition combinations, individual incentives are not possible to be predicted.

To address these gaps in experimental studies, eight particular treatments which have unique equilibrium coalition are employed in this study. In these treatments, each agent has a weakly dominant strategy to follow. The individual preference is therefore identifiable and can be observed.

This design offers two main advantages: firstly, this study endeavours to investigate incentives for participating in IEAs. If a coalition has more than one equilibrium, individual decisions cannot be predicted. But, if we have a coalition with a unique equilibrium, it would provide a suitable environment to observe individual decisions when every player has a best strategy to make. Secondly, the hypothesis of this study assumes that the other-regarding preference would influence the equilibrium differently from the egoistic preference. This entails that a coalition would be formed differently when individuals care about others agents' payoffs.

To the best of our knowledge, what motivates individuals to participate in a public goods coalition has not yet been fully explored in the existing literature. This study asks the following questions: Does the concern about fairness change players' decisions? If so, how much would they care? How do individuals' social preferences affect their own incentives for participating in a public good game?

To answer these questions, we have designed an experiment as follows. It comprises of two parts: the first part aims to find out the individual inequality-averse preference. The subjects of the experiment are paired and asked to choose from a certain fair payoff and an all-or-nothing payoff. When the expected payoff is higher than the fair payoff, those who prefer to have the fair payoff would be considered as inequality-averse players. They would be more likely to break the internal and external constraints in the coalition game.

The second part is a public good game. The subjects are grouped into 5-player groups. Since our main interest lies in the formation of IEAs, the experiment has taken out the abatement game, and turned it into a public good game which mimics the membership decision process. The subjects are given particular payoff tables to decide whether or not to join the coalition. Bearing in mind the results from the first experiment, the predictions with the other-regarding preferences are expected to explain a smaller free-riding effect and various coalition combinations.

Our theoretical finding predicts that, if the internal and external constraints hold and the condition for the unique equilibrium is satisfied, the coalition formation could be either a unique $n^*$-member coalition, or a unique coalition which is larger than $n^*$, or an unstable coalition with different inequality-averse preferences. The constraints could be violated when agents have strong attitude of inequality-aversion. However, our experimental evidence does not fully support the theory. In terms of the individual decisions, when subjects could free-ride, those with a higher marginal benefit were less likely to join a coalition and prefer to have a lower payoff. On the other hand, the subjects with a high degree of religious belief were more likely to be free-riders by not joining a coalition and having higher payoff.

From the questionnaire in the experiment, we learnt that right-wingers are more likely to build a larger coalition when they could be free-riders. Comparing to the results on the internal constraint, right-wingers are more likely to violate both internal and external constraints. Right-wingers tend to act strategically by punishing and compromising when they are in different

roles.

The article is structured as follows. After the introduction, in Section 2, we build a model based on heterogeneous players. In order to investigate the individual behaviour in the coalition formation, we will focus on the case of unique equilibrium coalitions. In Section 3, the data from two experiments which are based on the theory discussed in Section 2 will be presented. In Section 4, we discuss the implications of the model and possible applications, and conclude. The theoretical proofs, the instructions of the experiment are included in the appendices.

## 2. The model

### *2.1. Benchmark model with heterogeneous players*

Supposed there are $N$ countries with different marginal benefit of total abatement, we label them as country 1, 2, ..., $N$. There are now $2^N-(N+1)$ possible coalition combinations [1]. In order to clarify, we assume that player 1, 2,..., $n$ are in the group to form an IEA, player $n+1$, $n+2$, ..., $N$ are not [2]. We rank $n$ countries in the coalition according to the value of their marginal benefit of abatement going from high to low as $\gamma_1 > ... > \gamma_n$. On the other hand, the nonsignatories are also ranked from high to low as $\gamma_{n+1} > ... > \gamma_N$ Any marginal benefit of total abatement ($\gamma_k$, $\forall k \in [1, ..., N]$) is in the range between 0 and 1[3]. The unit cost of abatement for each country is assumed as 1.

Each country faces a game which run in two stages: at the first stage, players play a membership game where they decide whether to participate in the coalition or not. At the second stage, given the decision made at the first stage, signatories and nonsignatories play the emissions abatement

[1] Any coalition needs at least 2 players. No coalition is a possible solution if no one cooperates.

[2] Any coalition needs at least 2 members, so $n \in [2, N]$.

[3] The meaningful range of the marginal benefit of total abatement ($\gamma_k$) is between 0 and 1. When $\gamma_k$ is too large ($1 \leq \gamma_k$), an IEA is unnecessary because players already have the incentive to abate fully. When the aggregate marginal benefit is too small ($\sum_{k=1}^{N} \gamma_k \leq 1$), a profitable IEA is also non-existent because all players would pollute anyway. When the marginal benefit is in between, there may exist stable coalitions where signatories abate and nonsignatories pollute.

game respectively. Each nonsignatory makes her own decision on emissions abatement with the objective of maximising her individual payoff. Meanwhile, members follow a common decision on abatement with the common objective of maximising the coalition payoff. We solve this two-stage game by backward induction.

We start with the abatement game. Let any nonsignatory $j$'s abatement be denoted by $x_j$. In order to simplify the model, the cost and benefit functions are both linear and the normalised level of abatement ($x_j$) is in the range between 0 (implies full pollute) and 1 (implies full abate).

With a profitable $n$-member coalition, a nonsignatory $j$'s payoff function $\pi_j$ with respect to the abatement $x_j$ is presented as

$$\begin{aligned} \max_{x_j} \pi_j &= (-x_j) + \gamma_j X \quad \forall \text{ nonsignatory } j = n+1, ..., N \qquad (1) \\ \text{where } X &= \sum_{i=1}^{n} x_s + \sum_{j=n+1}^{N} x_j \end{aligned}$$

where $x_j$ is the individual abatement with its marginal benefit rate $\gamma_j$ [4]. $X$ is the total abatement which includes $n$ signatories' aggregate reduction ($\sum_{i=1}^{n} x_s$) [5] and $(N-n)$ nonsignatories' aggregate reduction ($\sum_{j=n+1}^{N} x_j$). From the first order condition of (1) with respect to $x_j$, we have polluting ($x_j = 0$) is the best strategy for a nonsignatory $j$.

For any signatory $i$, all members act as one to maximise the coalition payoff and share this coalition payoff equally. The $n$-member coalition payoff ($\Pi^s$) is the overall payoff of all members ($\pi_i$). The coalition payoff is maximised by choosing the common abatement ($x_s$)

$$\max_{x_s} \Pi^s = \sum_i \pi_i = \sum_i^n [(-x_s) + \gamma_i X] \qquad (2)$$

[4] $\gamma_j \in \{\gamma_{n+1}, ..., \gamma_N\}$

[5] Because members in the coalition move as one, the aggregate emission abatement would be $\sum_{i=1}^{n} x_s = n \cdot x_s$.

From the first order condition of (2) with respect to $x_s$, we have

$$\frac{\partial \Pi^s}{\partial x_s} = -n + n \sum_i^n \gamma_i = 0 \tag{3}$$

When $\sum_i^n \gamma_i < 1$, polluting is the best strategy but then the coalition would be meaningless. To form a profitable coalition, the total contribution should go beyond the threshold which the sum of marginal benefit of members is lager than 1 ($\sum_i^n \gamma_i \geq 1$) and the best strategy for all members is fully abating ($x_s = 1$).

Since the coalition aims to maximise its payoff, individual decisions of members should achieve this goal. Burger and Kolstad (2009) note that majority voting rule, unanimity and joint payoff maximisation are all equivalent under the assumption of homogeneous agents. However, with heterogeneous agents, they suggest that majority voting reflects the interests of the median voter and may not reach a joint payoff maximum. Although wealth transfers among member of the coalition is often suggested as being politically infeasible, Kolstad (2014) states that "sharing the wealth" within the coalition might be appropriate.

Hence, to achieve the goal of maximum a coalition payoff, each member shares the same responsibility. We assume that the coalition payoff is equally shared by all signatories. Any signatory $i$ with a $n$-member coalition has a payoff

$$\pi_i = \frac{1}{n} \Pi^s \tag{4}$$

It should be noted that a rule of the coalition requires coalition members using transfers to equalise net payoffs between agents. Such rule achieves a less unequal distribution of payoffs through transferring. This assumption implies that for the main purpose of this study, it is difficult to separate out the issue of IEA formation and its impact on fairness from the fact that the IEA is itself a mechanism for achieving a less unequal distribution of payoffs through using transfers. Countries with higher marginal benefit of the total abatement are more likely to leave the coalition *ex post*, because those countries could earn higher payoff for the absence. However, we assume that countries have the full information when they agree to participate in an IEA, they know the consequence of being signatories and nonsignatories. Signatories will commit to stay in the coalition and make transfer to equalise

individual payoffs. We appreciate that this is a strong assumption[6]. However, considering each member have to move as one to maximise the coalition payoff, every member would share equally responsibility. Hence, our design of sharing the coalition is still an adequate solution.

Hence, the payoff of a signatory $i$ in a profitable coalition is

$$\pi_i = -1 + \sum_{i}^{n} \gamma_i \tag{5}$$

In the membership game, players are asked to decide to participate in a coalition or not. The decisions are made simultaneously. With the internal and the external constraints by D'Aspremont *et al.* (1983),

$$\text{Internal constraint} \quad : \quad \pi_n^s(n^*) > \pi_n^{ns}(n^* - 1) \tag{6}$$

$$\text{External constraint} \quad : \quad \pi_N^s(n^* + 1) < \pi_N^{ns}(n^*) \tag{7}$$

There exist stable coalitions. The *internal constraint* (6) denotes that a signatory has no incentive to leave the $n^*$-member coalition and $n^*$ is the stable number to maintain the coalition. If it is satisfied, every one would like to participate in the coalition. The *external constraint* (7) describes that a nonsignatory has no incentives to participate in a coalition as the $(n^* + 1)$-th member. If it is satisfied, all nonsignatories do not want to participate [7].

This study attempts to test the theory based on heterogeneous agents by conducting an experiment. Existing experimental studies (such as Kosfeld *et*

[6]The rule would deter a country to abandon its commitment on membership by some policies, e.g. high penalty punishment and international sanction.

The issue of different policy instruments of transfer and commitment could be discussed by further studies.

[7]The stability of the coalition can be explained with two 3-player cases. In case (i), if the aggregate marginal benefit of total abatement is too small to form a profitable coalition, there is no stable IEA. For example, when the set of the marginal benefit of players 1, 2 and 3 is $\{0.4, 0.3, 0.2\}$, no player would like to participate because all possible combination are unprofitable.

In case (ii), when the aggregate marginal benefit is high enough, there might exist an equilibrium or equilibria coalitions. For example, given the set of marginal benefit is $\{0.7, 0.6, 0.35\}$, there exist two stable coalitions $\{1, 2\}$ and $\{1, 3\}$. In the former case, the internal constraint is satisfied when both players 1 and 2 have no incentive to dissolve the coalition by leaving. On the other hand, the external constraint is satisfied when player 3 has no incentive to join since the reward of free-riding is better than that of participation.

*al.*, 2009) assume that all agents are identical. However, this assumption is far from the reality. Even the assumption of heterogeneity is considered by Burger and Kolstad (2009), there exist more than one equilibrium coalition in their experimental design. Though the formation of IEAs could be expected, it is not enough to predict individual decisions in the membership game by these past studies. In order to address this gap in the literature, this study considers the condition of uniqueness of equilibrium. The condition provides the existence of a unique stable $n^*$-member coalition where $n^*$ is the minimum number to form a profitable coalition. By this condition, individual decisions could be predicted.

**Condition 1.** *(Uniqueness of equilibrium)*
*Suppose all players are self-interested, when the internal and the external constraints are satisfied, there exists a unique stable $n^*$-member coalition if and only if* $1 + \gamma_{n^*} > \sum_{i=1}^{N} \gamma_i$

The proof is presented in Appendix 1.

The condition implies that the stable coalition is unique if the absence of any single signatory cannot be replaced by the entry of all nonsignatories. The unique equilibrium condition ensure that the formation is the only one profitable coalition ($-1 + \sum_{i=1}^{n^*} \gamma_i > 0$). If any player from player 1 to $n^*$ leaves the coalition, there is no substitution to form a profitable coalition. Connecting the internal constraint ($\sum_{i=1}^{n^*} \gamma_i > 1$) with the unique equilibrium condition, we have

$$\sum_{i=1}^{n^*} \gamma_i > 1 > \sum_{i=1}^{n^*-1} \gamma_i + \sum_{j=n^*+1}^{N} \gamma_j$$

By subtracting $\sum_{i=1}^{n^*-1} \gamma_i$ from both sides, we derive that

$$\gamma_{n^*} > \sum_{j=n^*+1}^{N} \gamma_j$$

Whilst we acknowledge this indeed a strong condition, however, in order to identify the individual incentives to participate in the coalition, such a condition provides an environment where each agent has a weakly dominant strategy in terms of their own payoffs.

*2.2. Inequality-averse preference in a coalition game*

The constraints above are considered assuming individuals have egoistic preferences. As mentioned previously, this assumption fails to capture the idea that individuals may behave differently in a practical interactive game. In order to address this limitation, we now incorporate the idea of "other-regarding" preferences into our analysis to examine individual incentives.

Following Fehr and Schmidt (1999), we assume that subjects dislike unfair outcomes at different levels. Subjects feel disadvantaged when they are better off or worse off in material terms. With this concept, the utility of a player $k$ of a profitable $n$-member coalition can be represented as

$$\begin{aligned} & u_k(n) \qquad (8) \\ = \; & \pi_k(n) - \frac{\alpha_k}{N-1}\sum_{k' \neq k} \max\left(\pi_{k'}(n) - \pi_k(n), 0\right) - \frac{\beta_k}{N-1}\sum_{k' \neq k} \max\left(\pi_k(n) - \pi_{k'}(n), 0\right) \end{aligned}$$

where Player $k'$ denotes all players except player $k$. The first term is the payoff of player $k$ and the second term indicates the average utility loss from other player $k'$ with the disadvantage-loss parameter $\alpha_k$. The third term measures the average loss from other player $k'$ with the advantage-loss parameter $\beta_k$, which is assumed within the range between 0 (inequality-neutral) and 1 (highest degree of inequality-aversion).

When the individual inequality-averse preference is considered in the utility function, the internal and the external constraints (6) and (7) and be rewritten as

$$\text{Internal constraint} \quad : \quad u_n^s(n^*) > u_n^{ns}(n^* - 1) \qquad (9)$$

$$\text{External constraint} \quad : \quad u_N^s(n^* + 1) < u_N^{ns}(n^*) \qquad (10)$$

An unique $n^*$-member coalition will still exist when all agents are self-interested. When an agent has a high attitude of individual inequality-aversion, the internal and the external constraints (9) and (10) could be violated. We have the following hypothesis on the coalition formation.

**Conjecture 2.**

*If the internal and external constraints hold and the condition for the unique equilibrium is satisfied, the coalition formation could be either a unique $n^*$-member coalition, or a unique coalition which is larger than $n^*$, or an unstable coalition with different inequality-averse preferences.*

The explanations of the possible outcomes are shown in Appendix 2. Three possible outcomes are depending upon different circumstances of individual inequality-averse preferences :

(i) When all players have no inequality-aversion or a low degree of inequality-aversion, there exists a unique $n^*$-member coalition equilibrium.

(ii) When any player from players $n^* + 1$ to $N$ has a high degree of inequality-aversion (large $\beta$), the external constraint could be violated. If other things are equal, the coalition formation is stable and larger than $n^*$.

(iii) When any player from players 1 to $n^*$ has a high degree of inequality-aversion, the internal constraint could be violated. The coalition formation then becomes unstable.

Without taking inequality-aversion into account, a unique stable coalition is formed with three constraints. When the inequality-aversion is considered as part of the individual preferences, there are a number of effects. First, inequality-aversion reduces countries' utility when payoffs are not equal. The incentive of being a nonsignatory therefore decreases and the external constraint is more likely to be violated. This will tend to increase the size of a stable coalition.

Second, countries with strong inequality aversion would be encouraged to stay in an IEA or join it to spread the benefits of equalisation because of the transfer mechanism where signatories share the same coalition payoff. However, except for a grand coalition, any combinations of IEAs has a free-riding effect. An expanding IEA will tend to exacerbate the payoff gap between signatories and nonsignatories. Signatories with a strong sense of inequality-aversion may violate the internal constraint if the payoff gap is large. Under this condition, the most likely outcome would be to have no IEA at all, so a certain level of inequality aversion can destabilise an IEA.

When inequality-aversion is taken into account, the net effect of these two factors shapes the stability and the formation of IEA. When a country decides to join a coalition given the first effect, the participation will lead to a smaller advantage loss but a larger disadvantage loss. With this character, a stable coalition can not be easily expand by the first effect. On the other hand, as long as stable equilibrium is not a grand coalition, there exists inequality. The payoff gaps between signatories and nonsignatories are enlarged with the second effect. The internal constraint is more difficult to be satisfied and the coalition formation becomes unstable.

In terms of the design of this particular example, the external constraint will not be violated given the highest degree of inequality-aversion. Hence, a

larger stable coalition is not possible in this case.

## 3. Experiment design and procedure

The experiment was conducted at the centre for EXperimental EConomics (EXEC) laboratory at the University of York (UK) and programmed with z-Tree (Fischbacher, 2007). There were 50 subjects who were registered on the ORSEE registration system by Greiner (2004). They were students from different countries and in various disciplines at the University. This sample that mimics the diversity in the real world where international policy makers and multidisciplinary knowledge are present helps understand IEAs formation. To ensure the data quality, the subjects had to comprehend the rules of the game as much as possible. To do so, the experimenter introduced the rules and gave the participants time to read through the instructions thoroughly and accomplish the controlled questions. In the end of each part of the experiment, four control questions were asked to test the subjects' understanding of the payoff tables. A new part would only start if all subjects had answered all control questions correctly.

A questionnaire was circulated before the experiment to gather demographical information, including the subject's degree disciplines, age (the year they were born), ethnicity, political orientation, and the level of belief in a religion. This questionnaire is designed to gather more explanation on their decision-making in the experiment. The first three questions are objective and the data shows the diversity of the participants. The distribution of subjects' major are: 11 participants recruited were reading Economics; 8 participants in Humanities; 13 participants in Science; 1 participant in Laws; 9 participants in Engineering; 1 participant in Psychology; 7 participants in other disciplines and no recruit was reading Business-related disciplines. The distribution of ethnicity: 32 subjects were white; 15 were Asian or Asian British; 2 were Black or African or Caribbean or Black British; and 1 fell into the category of any other ethnic groups. Also, all participants were undertaking undergraduate or postgraduate courses at the University and their average age was 25 years-old (the oldest being 45 and the youngest being 21).

The last two questions were concerning their subjective preferences. The distribution of their level of belief on religions while subjects were asked to identify themselves on a scale ranging from level 1 (not religious at all) to 5 (extremely religious). In the results, 20 subjects consider themselves to be

atheist. Meanwhile, 6, 8, 9, and 7 subjects consider themselves as religious, with mild belief, median belief, strong belief and pure religionists respectively. The average level is 2.5. The distribution shows that the subjects' religious belief is between mild to median belief, overall. The other question aims to indicate the subjects' political preference (level one indicates left, level two centre-left, level three neutral, level four centre-right and level 5 right). In our sample, 7 subjects self-identified themselves as left wing; 10 as centre-left; 25 as neutral; 7 as centre-right and 1 as right wing.

The experimental procedure was designed as follows.

*3.1. An inequality-averse preference test*

This test aimed to examine the subjects' individual attitude towards inequality-aversion. To measure a subject's inequality-averse preference, the subjects who did not know each other were paired together. **The subjects did not know their partners and the partners' decisions during the whole test.** Their payoffs were determined by their own decisions as well as their partner's decisions. This was to understand the individual preferences without knowing their strategies they played. The subjects were required to answer a series of decision questions in 11 rounds as shown in Table 1. Option 1 meant the subjects share the same allowance, while Option 2 meant the subjects could take all-or-nothing with a certain probability.

Given the allowance £5, which would be shared by a subject (denoted as $A$ afterward) receiving $x$ and another subject (denoted as $B$ afterward) receiving $(5-x)$. Subject $A$'s inequality-averse utility was determined by both her and the other subject's shares as displayed in Table 1.

$$U_A(x, 5-x) = \begin{cases} x - \alpha[(5-x) - x] & \text{if } x \leq 2.5 \\ x - \beta[x - (5-x)] & \text{if } x \geq 2.5 \end{cases} \tag{11}$$

The upper function represents Subject $A$'s utility when $A$ has less than half of the total allowance. The parameter $\alpha$ is the coefficient of the average disadvantage loss of $A$. On the other hand, when $A$ has more than half of the total allowance, the lower function is $A$'s utility with the coefficient of the average advantage loss.

The function can be presented as the solid line in Figure 1. The horizontal axis is the allowance of $A$ while the vertical axis is $A$'s corresponding utility.

The utility depends on the payoff set of subject $A$ and the opponent $B$ which is presented as (£$x$, £$5-x$). From (11), we derive that $A$ 's utility of

(£5, £0) is $U_A(5,0) = 5 - 5\beta$, and the utility of (£0, £5) is $U_A(0,5) = -5\alpha$, and the utility of (£2.5, £2.5) without any inequality is $U_A(2.5, 2.5) = 2.5$. We normalise by setting $[U_A(5,0) - U_A(0,5)]/5 \equiv 1$.

Given that a series of probabilities is involved in the inequality test, this test could be characterised by strategic uncertainty. The subjects' risk attitudes may be involved in their decisions. For instance, even the expected payoff of taking Option 2 is higher than the payoff of Option 1, a risk averse subject may prefer to the equal-share option because she or he fears the possible loss by taking Option 2. There are some experimental designs, such as Blanco *et al.* (2011) and Yang *et al.* (2012), that attempted to exclude strategic uncertainty. and avoid risk attitudes. They employed two games to capture the factors that advantage or the disadvantage the subjects.

The relationship between risk-aversion and inequality-aversion has been discussed by several recent studies. An experimental study by Carlsson *et al.* (2005) also found that people who are inequality-averse are more risk-averse, and that the reverse relation also holds true: risk-averse individuals tend to be more inequality-averse. Given the same individual risk, Kroll and Davidovitz (2003) provided another experimental evidence that most of the subjects preferred equal distribution to inequality.

Whilst it should be noted that our experimental design did not exclude the subjects' risk attitudes, our design is still superior in the sense that the normalisation provides a normalised inequality-averse utility in one game[8]. While other studies avoid strategic uncertainty in their experiments, there exist other factors which could lead to a biased estimation of inequality-aversion. For example, Yang *et al.* (2012)'s experiment shows that subjects may have a negative advantage loss. It implies that subjects may prefer to show off rather than feel guilty when they are advantaged. Such bias does not arise in our design because the utility has been normalised.

To find out the inequality-averse preference, we asked each subject to choose between two options in each row of Table 1. The first option is a certain option where both players share the allowance equally (£2.5). The second option is an uncertain option that the subject would win all-or-nothing depending on probability. The given probability decreased by 10% in each

[8]We acknowledge that there are other methods to measure attitudes to inequality. Different to other experiments focus on social preferences, there were two social preferences tests and one public good game in our experiment. This design could measure individual inequality-averse attitude without complex procedures.

round.

Since the allowance was a good, the subjects in theory would prefer to have more. The first row in Option 2 shows that if the probability to yield (£5) is 1, any subject would choose Option 2 rather than Option 1. On the other hand, at the bottom row in Option 2, if the probability of the set (£5, £0) is equal to 0, subjects would prefer Option 1 rather than Option 2. Hence, we assume that subjects will choose Option 2 in the first few rows and Option 1 in the last few. For each subject with a consistent preference, there exists a point with a certain probability where the subject would switch from Option 2 to Option 1. We denote the probability of (£5, £0) at the switch point by $p$. Then subjects feel indifferent between (£2.5, £2.5) for sure and (£0,£5) with probability $(1-p)$ and (£5, £0) with probability $(p)$. Such probability $p$ can be seen as the weight of inequality aversion.

In Option 2, a subject is given (£5) with the probability $p$ and (£0) with the probability $(1-p)$. In Option 1, the subject is given (£2.5) for sure. The subject would feel indifferent between the sharing combination (£2.5, £2.5) and the mixed combination of (£0, £5) with probability $(1-p)$ and (£5, £0) with probability $(p)$. We can present this in an equation as

$$U\left(2.5, 2.5\right) = \left(1-p\right) U\left(0, 5\right) + pU\left(5, 0\right) \tag{12}$$

The inequality-averse parameters $\alpha$ and $\beta$ would be found through $p$. Given that the range between the utility of all $U\left(5, 0\right)$ and nothing $U\left(0, 5\right)$ is normalised, the inequality-averse preference was indifferent when subjects are disadvantaged and advantaged ($\beta = \alpha$). Although it was mentioned earlier that a player might suffer more from inequality when she is disadvantaged ($\beta \leq \alpha$), there are two reasons that support us to do so. In practice, it is not easy to find a subject's preference without standardising the unit of the utility. In the literature, the experimental evidences show that the disadvantage factor is not necessarily smaller than the advantage factor (Dannenberg *et al.*, 2007; and Yang *et al.*, 2012).

Hence, we assume that the inequality-averse preference are indifferent to being disadvantaged and advantaged.

When the subject is advantaged, $U(5,0)/U(2.5,2.5) = 1/p$, we have

$$\alpha = \beta = p - \frac{1}{2} \tag{13}$$

Since the probability $p$ is in the range of 0 and 1, the inequality-averse parameters $\alpha$ and $\beta$ are at the range of $-\frac{1}{2}$ to $\frac{1}{2}$.

Subjects are inequality-neutral when their switch points are at $p = 0.5$ where the expected payoff is equal to the fair payoff. The inequality-averse preference $\alpha = \beta = 0$ . In other words, the utility of taking all the allowance (£5) is not two times higher than that of equally sharing the allowance (£2.5). When the switch point is $p > 0.5$, subjects are inequality averse and their utilities are lower than their monetary payoffs. The extreme case is when $p = 1$, and $\beta$ is 0.5. It implies that subjects have indifferent preferences of taking one unit payoff or equally sharing the allowance. When the advantage aversion is very high ($\beta > 0.5$), it is considered as altruism, which is not able to capture in this design[9]. Altruists would prefer to give goods to others in order to achieve fairness. Although it is beyond the scope of this study, altruism is an important topic that needs to be explored in future studies as it can happen in reality.

When $p$ is less than 0.5, subjects are not inequality-averse (neither advantage acceptors nor disadvantage acceptors). While Fehr and Schmidt (1999) exclude inequality acceptors in their assumption, inequality-aversion is considered in this study as it may happen in the experiment. For these subjects, they would be considered as inequality-lovers or risk-lovers (because the experiment has strategic uncertainty). Both inequality and risk lovers are possible but uncommon in reality (as seen in the experimental result later), so our study does not focus on this issue. Hence, these subjects have been excluded from our sample[10].

*3.2. Experiment of a coalition game*

To concentrate on the entry decision, we simplify the two-stage game into the membership game. The game has been modified to show the situation when a *profitable* $n$-member coalition is formed (the coalition generates a positive payoff if the aggregate benefit-to-cost ratio of signatories is larger than 1, $\sum_i^n \gamma_i > 1$). In this case, all signatories abate and all nonsignatories pollute. Otherwise, the coalition *collapses* and all players pollute. Hence, all elements in the payoff set $(\pi_1(n), \pi_2(n), ..., \pi_N(n))$ are non-negative. It implies that all players pursue their self-interest in maximising their own payoffs.

[9]Because the probability $p$ is only in the range of 0 and 1.

[10]The existing probabilities in the test may introduce a bias by involving risk-averse preference and hence weaken the thoery by .

The social welfare is the aggregated payoffs from all nonsignatories and the coalition payoff . The maximum welfare exists when the grand coalition is formed [11]. All players face a dilemma of being a nonsignatory with free-rider payoff or being a member with the shared payoff.

A public good game with various payoff tables was conducted. The results from the previous part were used to predict whether the subjects would violate the stability constraints in the coalition game. In the theoretical model, this is a two-stage game. The first stage is the membership game, where subjects decide whether or not to join a coalition. The second stage is the abatement game. In the abatement game, since the payoff is a linear function, the decision-making would be straightforward. When a subject decides to join a coalition, she would abate fully at the second stage. When her decision is not to join, she would not abate at all at the second stage. Based on this, we simplify the two-stage model to a one-stage membership game in the experiment.

In this part, subjects were randomly assigned to groups of five persons. They did not know who they were playing with, but they did know that they were playing with the same people during the whole session. In our assumption, subjects should be self-motivated. Subjects were therefore required to maximise their own payoffs.

In each treatment, each subject was given a particular payoff table of all the possible coalition combinations. A group of $N$ subjects would generate $\left(2^N - N - 1\right)$ combinations. In order to generate a simple and clear table for subjects, the number of 5 subjects was set in a group with 26 possible combinations.

The game was a one-shot game, and decisions in each round were independent. With this design, the subjects know no more than their own inequality-averse preference. However, the experiment allowed subjects to have a learning process so that the coalition would converge to the Nash equilibrium. The game was played 15 times in each sub-treatment. Subjects were given 180 seconds to make their decisions of whether or not to join the coalition. According to the pilot experiment, this time setting gave subjects enough time to make their decisions. Any decision which was not made

[11] The total payoff is $\Pi = \Pi_s + \sum_j \pi_j = [(-n) + \sum_i^n \gamma_i] + \left[n \sum_j \gamma_j\right]$. Because only a profitable coalition is counted, the total payoff is maximised when the grand coalition is formed $\Pi = (-N) + \sum_i^N \gamma_i$.

within this amount of time would be counted as non-participation. This rule is sensible because the decision was asked whether or not to join a coalition with a non-participating status.

Finally, the coalition formation and all subjects payoffs in the group were reported on the result screen.

When subjects are self-motivated, they therefore maximise their own payoffs. Subjects should make their decisions based on their economic incentive. In order to ensure subjects were aware of their profit-maximising incentives rather than other non-economic incentives, the reference to environmental issues was removed from the instruction. The level of marginal benefit of the total abatement was labelled as parameter ($\gamma_k$, $\forall k \in [1, ..., 5]$) in the experimental design. There are two treatments with different parameter sets. 20 subjects took Treatment 1 and the rest of the subjects took Treatment 2. The individual parameters in the Treatment 1 are listed in Table 2, and the parameters in Treatment 2 are listed in Table 3.

According to Condition 1, we can claim that a unique equilibrium could be found in some particular cases. The theoretical result suggests that a unique equilibrium exists within the internal, the external and the unique constraints. To achieve a unique equilibrium, the experiment was built with some particular parameters mentioned earlier in the theory. Subjects with high marginal benefit parameter are labelled ($^*$) in Tables 2 and 3, they were predicted to have a weakly dominant strategy to *join*. Eight treatments within the constraints were selected in the experiment. The theoretical size of the stable coalition in treatments was from 2 to 4. Each group was given four sub-treatments with a different number of subjects predicted to be in the stable coalition.

Tables 2 and 3 present the treatments which were designed to ensure a unique stable IEA based on the assumption of no inequality-aversion. Each sub-treatment had a unique equilibrium and each subject had a weakly dominant strategy in the membership game. Meanwhile, we propose in Conjecture 2 that different attitude to inequality-aversion may lead to higher membership or no stable IEA. The internal constraint is more likely to be violated by individuals with high degree of inequality-aversion. But due to the internal transfers, a nonsignatory would gain less advantage loss but more disadvantage loss if she or he decides to join a coalition. Hence, the external constraint is not easy to be violated. The experiment in this study is able to test whether subjects with high inequality-aversion are more likely to violate

the internal constraint and lead to unstable.

When a possible coalition is unprofitable, all subjects in the group gain nothing for return. The possible payoffs for subjects were from £0 and up to £24. The payoff depended on the given parameters and the coalition formation. In the experiment, we simplified the decision-making process by reducing the calculation process. With the payoff table, subjects could easily find the corresponding possible payoffs without working on the payoff function.

In the case of the external constraint, we assume that all subjects are inequality neutral except for Player 1. Player 1 would obey the constraint if the utility of being a nonsignatory $(6.75 - \frac{2.25}{4}\alpha - \frac{15.75}{4}\beta)$ is better than being a signatory $(3.75 - \frac{8.25}{4}\alpha)$ . However, the subject would violate the external constraint when she has high inequality aversion. Since the disadvantage-aversion is indifferent to the advantage-aversion, Player 1 would violate the external constraint when $\frac{16}{13} < \alpha$ (or $p > \frac{45}{26}$) . However, altruism cannot be captured in this test because Player 1 is unlikely to join the coalition with Players 3, 4 and 5, as mentioned earlier.

Similarly, Player 2 would violate the external constraint only when the subject's preference $p > \frac{37}{26}$. It means that Player 2 is very unlikely to join the coalition.

In the case of the internal constraint, if others are inequality-neutral, Player 3 would follow the internal constraint when the utility of joining $(1 - \frac{8.5}{4}\alpha)$ is higher than the utility of not joining (0). However, if Player 3 has strongly inequality-averse preference, $p > 0.97$, Player 3 would violate the internal constraint and not join the coalition. With the unique coalition condition, whether the external constraint is obeyed by others or not, the equilibrium would be a failed coalition because Players 3, 4 and 5 are irreplaceable.

Similarly, Players 4 and 5 would violate the internal constraint if their preference $p > 0.97$.

We can therefore calculate the threshold to break the internal and external constraints for each subject. Subjects who break the external constraint would have very high advantage aversion. However, we should note that altruism can not be captured in our test. On the other hand, the internal constraint is more likely to be violated. The thresholds are also very high. This could explain that subjects are likely to follow their weakly dominant strategies.

*3.3. The results from the experiment*

In the inequality-averse test, each subject was asked to choose from two options in 11 rounds. In the theoretical prediction, the decision in round 1 would be 'Option 2' and the decision in round 11 would be 'Option 1'. One turning point was expected and that was when the decision changed from Option 2 to Option 1. The result demonstrates that 33 out of 50 subjects had no more than one switching point in 11 rounds, while 2 subjects took Option 1 in the whole part. The degrees of inequality-aversion were therefore determined.

Figure 2 presents the number of subjects taking Option 1 in each round. The majority had their switch point at rounds 3, 4, 5, or 6. After round 7, almost every subject took Option 2. Although the experimental design allowed the existence of inequality acceptors, as predicted in the assumption of the theory, the degree of inequality-aversion was unlikely to be negative. As mentioned earlier, five subjects were excluded because they were negative inequality-averse.

Table 4 shows the OLS estimation of inequality-averse preference. The dependent variable is average times of taking the Option 1 in the inequality-averse test. Independent variables are subjects' age (AGE), political attitude (POLITIC), and religious attitude (RELIGION). The result shows that these factors from our questionnaire have insignificant effect on subjects' inequality-averse preferences.

In the membership game, all subjects were put into 10 groups and took four sub-treatments in 60 rounds. Groups 1 to 4 used Treatment 1 in Table 2 and groups 5 to 10 used Treatment 2 in Table 3. Each subject in the group was given a different value of the marginal benefit parameter $\gamma$. This parameter implied their contribution to the group, if they decided to join in. When the total contribution of a group was over 1, the coalition was profitable and everyone received the payoff which depended on their decisions. Otherwise, an unprofitable coalition brought nothing to all the players in the group. With the assumption of no inequality-aversion, the peculiar design of this experiment leads to a unique equilibrium and the total contribution of this stable coalition is 1.05.

Figure 3 shows the results of the total contribution of groups 1 to 4. The charts in the first row present the total contribution of groups 1, 2, 3, and 4 in sub-treatment 1 respectively. Similarly, the charts in the second, third and fourth rows present the total contribution of groups 1, 2, 3, and 4 in sub-treatments 2, 3 and 4 respectively.

Figure 4 shows the results of the total contribution of groups 5 to 10. The charts in the first row present the total contribution of groups 5, 6, 7, 8, 9 and 10 in sub-treatment 1 respectively. Similarly, the charts in the second, third and fourth rows present the total contribution of groups 5, 6, 7, 8, 9 and 10 in sub-treatments 2, 3 and 4 respectively.

In light of the study population, profitable coalitions were formed in 387 of 600 rounds. The various forms of group formation lead to different group payoffs. For example, group 6 and group 8 both take Treatment 2. Group 6 forms profitable coalitions in 47 rounds, but group 8 achieved that in only 12 rounds. Both treatments provided subjects with weakly dominant strategies to take. If subjects in the group all made their weakly dominant strategies, the internal and external constraints were held and the coalition was at Nash equilibrium. It happened in 112 out of 600 rounds and such a coalition was not stable as predicted in the theory. According to the experimental results, more than two third of the profitable coalitions were formed and they were larger than the Nash equilibrium size.

In order to test our hypothesis, we examine the subjects' decision in the past round and their individual inequality-averse preferences to predict their next move. The indicated level of inequality-aversion is therefore employed to predict individual decisions in a coalition game. Figure 5 presents the total contribution of Groups 1 to 4. Similarly, the actual total contribution and the predicted total contribution with and without inequality-aversion of Groups 5 to 10 are shown in Figures 6 and 7.

The blue line with spots in each chart presents the actual total contribution in a sub-treatment. Given the results in the past round, the red line with cross marks are the prediction of the total contribution with the decision in the past round and subjects' individual inequality-averse preferences. There are two main reasons for employing this prediction. First, the subjects know their own inequality aversion parameter, but not others. The test in Part 1 of the experiment was anonymous and independent of Part 3, the subjects should not learn others' inequality-averse preferences. Second, learning and reciprocity are not considered in our model. Though the experiment design allows subjects finding their dominant strategy, it is not expected to figure out other's social preference. Since the subjects know no more than their own individual preferences and the historical decisions on the membership

game, our prediction should be based on such information[12].

In order to examine our conjecture, the green line with triangle marks is generated only with the individual decisions in the past round only. In other words, this predictions are based on the assumption of neutral inequality-averse preference.

Compared to these neutral predictions, in most cases, the predicted total contributions with inequality-aversion is higher. Both predictions are higher than the profitable threshold during the whole experiment. When subjects have high inequality-aversion, the result is not as unprofitable as we expected. Besides, when the inequality-aversion is not taken into account, the predictions are more stable and closer to the actual outcomes.

When we examine the individual decisions, the predictions with inequality-aversion match the actual decisions by 1838 over 2800 observations (65.6%) while those neutral predictions match the actual decision by 74%. The internal constraint was not supposed to be violated but the results suggest otherwise. In the sample of 1540 observations, the predictions with inequality-aversion match the actual outcome at 77.2% of the observations, while those neutral predictions matched by 84.9%. On the other hand, the predictions on those observations when subjects should follow the external constraint are lower. Amongst the 1260 observations, the predictions with inequality-aversion matched by 51.5% and the neutral predictions match by 61.0%.

To further the discussion, the possible factors are examined by Maximum Likelihood Estimation(MLE) of binary probit regressions. The variables in Table 5 are the decision made at the last round (DECISION$(-1)$), the average number taking Option 1 in the inequality-averse test (INEQ), the year subjects were born (AGE), the political preference from left (1) to right (5) (POLITIC), the religion preference from atheist (1) to religionist (5) (RELIGION), the weakly dominant strategy from not joining (0) to joining (1) (WD STRATEGY), the marginal benefit of the total contribution ($\gamma$), and the total contribution of the group at the last round (TC $(-1)$).

As mentioned earlier, the data of five subjects has been excluded because their attitude to inequality is opposite to our assumption which says the sub-

[12]This experimental design attempts to purify the individual decision, any bias from other subjects' preferences should be minimised. It would be a potentially interesting but very complex issue to model (essentially testing Bayesian learning), we will leave this challenge to the future studies.

jects dislike inequality. We examine 45 subjects who have different degrees of inequality-aversion. The estimation of Probit MLEs(1) covers all observations of 2700 decisions which were made individually. Because two variables depend on the outcomes at the last round, only 2520 observations are used for the regression. Amongst these 2520 observations, the subjects decided to join 1692 times and not to join 828 times.

The inequality-averse factor (INEQ), the weakly dominant strategies (WD STRATEGY) and the decision at the last round (DECISION$(-1)$) have a positive effect on the decision at the 1% significance level. This interesting result implies that the higher inequality-aversion a subject has, the higher incentive this subject has to participate in the coalition. Also, when the decision at the last round or the weakly dominant strategy is being made, the subjects are more likely to choose joining. The marginal benefit of total contribution ($\gamma$) has a negative effect on decision-making at the 1% significance level due to the free-riding effect when the subjects' weakly dominant strategy was not to join. Nevertheless, it is insignificant even if the subjects join a coalition in the case where the total contribution at the last round (TC$(-1)$) is to join. Reviewing the factors listed in the questionnaire, (AGE) and (POLITIC) appear to be statistically insignificant. But, (RELIGION) has a negative effect at the 5% significance level. That means, the more religious a player is, the less likely s/he will join.

It was assumed that the subjects with a higher degree of inequality-aversion were more likely to violate the internal and the external constraints. In order to assess the internal constraint, we use Probit MLE(2) to examine the observations where the subjects' weakly dominant strategy was to join. 85% out of the 1500 observations obeyed the internal constraint. In our hypothesis, the subjects with a higher degree of inequality-aversion were expected to violate the internal constraint, and the coefficient of INEQ should be negative. However, interestingly, the results show that INEQ has a positive effect at the 1% significance level. This striking outcome implies that subjects with a higher degree of inequality-aversion are more likely to to join a coalition. Consequently, this outcome suggests that these subjects with a higher degree of inequality-aversion are less likely to violate the internal constraint. That said, the subjects have stronger incentives to form a profitable coalition when their sense of inequality-aversion is higher. Perhaps due to those subjects's preference of having a fair outcome, a safe act which could keep their payoffs low appears to be more favourable than a risky strategy of punishing other outsiders and forcing them to participate. Those with a

lower degree of inequality-aversion tend to act strategically. They usually attempt to punish free-riders from time to time and force outsiders to participate in a coalition. Such strategic behaviour makes the coalition process unstable over rounds. Comparing the experimental outcomes with the numerical example, we have observed instability in the coalition formation in the experimental results. The experimental results show that the instability is caused by the subjects with low degrees of inequality-aversion rather than those with high degrees.

The estimation of Probit MLE(3) tests the factors included in the questionnaire and the previous results. 1400 observations were collected, except for those in the first round where each sub-treatment was with weakly dominant strategies of joining. The internal constraint was violated 215 times. The result also supports a significant positive effect on the decision-making at the last round. The effect of (RELIGION) is rather insignificant in this test and (POLITIC) instead has a negative effect at the 5% significance level. It suggests that the pro-right-wingers violating the internal constraint is higher than that of the pro-left-wingers

This result could be explained in the example of Group 9. Four out of five subjects in the group had a switch point in the inequality-averse test. For example, Subject 44 had the highest degree of inequality-aversion - the switch point was at $p = 0.9$. The switch point of subjects 43, 45 and 41 were 0.8, 0.8, and 0.5 respectively. In the membership game, subject 44 violated the internal constraint in only three out of 45 rounds. The violation rates of subjects 43, 45 and 41 are 3%, 0%, and 43%. It shows that the subjects with a higher degree of inequality-aversion were less likely to violate the internal constraint.

However, the internal constraint could be broken by the subjects with a higher degree of inequality-aversion in a few cases. Group 5 where everyone in the group had a switch point in the inequality-averse test as a good example. Subject 21 had the highest degree of inequality-aversion and the switch point is at $p = 0.9$. Following that, the degree of subjects 22 and 24 is $p = 0.8$, the degree of subject 25 is $p = 0.7$, and subject 23 is inequality-neutral - the switch point is at $p = 0.5$. Subject 21 violates the internal constraint in 30% of the 30 rounds, while the violation rates of subjects 22, 24, and 25 are 13%, 0%, and 3% respectively. In this case, the subjects with a higher degree of inequality-aversion are more likely to act against the internal constraint.

The external constraint is assessed by the estimation of Probit MLE(4) where the observations' weakly dominant strategy is not-to-join. The con-

straint was violated in 45% of the 1120 observations. When a coalition is unprofitable, it is indifferent whether to join or not. Hence, the subjects would make a random decision in the next round. This is the reason why the external constraint was violated in almost half of the observations.

When a profitable coalition was formed, 44% of the subjects would violate the external constraint in the next round. If we only look at those subjects with a higher degree of inequality-aversion (INEQ$>$ 0.8), only 40% of them violated the constraint. Turning to the results from those with a low degree of inequality-aversion (INEQ$<$ 0.5), the constraint was violated in almost half of the observations. The result shows that the subjects with a high degree of inequality-aversion were more likely to be free-riders. This might appear to be counter-intuitive at first sight, but the subjects with a low degree of inequality-aversion have demonstrated different behaviour of forcing outsiders to participate when their dominant strategy was to join a coalition. When their roles changed to the opposite, they were more likely to compromise and cooperate.

The estimation of Probit MLE(5) examines the factors from the questionnaire. In our hypothesis, the marginal benefit of the total contribution ($\gamma$) has a significant negative effect on the decision. In contrast to the experimental evidence of Burger and Kolstad (2010), our results do not support their earlier finding that said that higher marginal benefits would significantly increase a coalition size and consequently the total contribution. This is mainly because our design limits any possible free-riding by excluding the subjects with high marginal benefit. This effect is shown in the estimation of Probit MLE(1). Despite the limitation of our design, the factor of the marginal benefit in the estimation of Probit MLE(5) is significantly negative and corresponds to the earlier findings. Our study provides more detailed information, compared to the existing literature, about how potential free-riding benefits would weaken the incentives for participation. When the dominant strategy is not to join, higher free-riding benefit comes with higher marginal benefit. The coalition size was likely to be larger than the equilibrium size when the players are with lower marginal benefits.

Our results can be summarised as below

**Summary 3.**

*In terms of the coalition formation, the predictions with inequality-aversion does not outperform those without.*

*In terms of the individual decisions when subjects could free-ride, those*

*with a higher marginal benefit were less likely to join a coalition and prefer to have a lower payoff. On the other hand, the subjects with a high degree of religious belief were more likely to be free-riders by not joining a coalition and having higher payoff.*

*Right-wingers are more likely to build a larger coalition when they could be free-riders. Comparing to the results on the internal constraint, right-wingers are more likely to violate both internal and external constraints. Right-wingers tend to act strategically by punishing and compromising when they are in different roles.*

## 4. Conclusion

This study has investigated the incentives to participate in IEAs with the other-regarding preferences, particularly the preference of inequality-aversion. The theory used in this study suggests that a stable coalition can be formed both internally and externally, when the signatories have no incentive to leave and the nonsignatories have no incentive to join. The assumption of inequality-averse preference argues that such a stable coalition would change by considering agents' preferences. Agents with a higher degree of inequality-aversion are more likely to break the internal constraint and leave the coalition.

A two-part experiment has been conducted to validate this theory. The first part was a test to measure the individual attitude to inequality-aversion. The second part was a public good game conducted to mimic the international environmental convention. Subjects were given different payoff tables and asked whether to join or not to join a coalition.

In order to fully capture individual behaviours in an IEA, the experiment has been designed in such a way that teased out as much noise and as many uncertainties as possible. In other words, the theoretical prediction for the experiment was purified to a unique equilibrium. In contrast to the existing literature, the results in this particular design do not support the theoretical prediction that a higher marginal benefit would significantly enlarge a coalition size and the total contribution. On the contrary, the subjects with a lower degree of inequality-aversion are more likely to act strategically by violating the internal constraint. By doing so, they could force free-riders to participate. But, when their role changes to the opposite, they reacted to compromise their payoffs.

Some other factors inquired in the questionnaire, such as the political preference and religion preference, have also shown significant effects on the decision-making. Pro-right-wingers behave as those with a lower degree of inequality-aversion and make more strategical decisions.

Although it is difficult to generalise solely based on one experiment which has its own limitations in design and data collection, this study has provided some promising results for understanding the real-world operation of IEAs, especially the dynamics that emerged during the decision making processes. One firm conclusion is that, in order to stabilise a coalition internally, international conventions had better emphasize the importance of fairness to signatories because a high degree of inequality-averse preference would lead a country to participate. An IEA could be enlarged when nonsignatories were informed of the potential damage if the target of the IEA cannot be achieved.

## Appendix 1

**Proof.** To proof the theorem, we establish an algorithm to find a stable coalition. Player $n^*$ has the incentive to maintain $n^*$-member coalition if the payoff $\pi^s_{n^*}(n^*) = (-1) + \sum_i^{n^*} \gamma_i$ is positive. If player $n^*$ leaves, the coalition collapsed. Hence, player $n^*$ gets $\pi^{ns}_n(n-1) = 1$ when all player pollute. When the internal constraint makes player $n^*$ to be stable in the coalition, all signatories have the same incentive to make it stable internally.

Meanwhile, the external constraint asks player $N$ to stay away from the $n^*$-member coalition. When player $N$ is a nonsignatory, its payoff is $\pi^{ns}_N(n^*) = (\gamma_N \cdot n^*)$. If player $N$ changes its mind and joins the coalition as the $(n^*+1)$-th member, the payoff becomes $\pi^s_N(n^*+1) = \left[(-1) + \sum_i^{n^*} \gamma_i\right] + \gamma_N$. When the external constraint deters player $N$ to join the coalition, all nonsignatories are deterred and the coalition becomes stable externally. Hence, the theorem is established.

By the internal and external constraints, the minimum number to form a profitable coalition is found. However, this coalition is not the only equilibrium. A coalition with more members could be another equilibrium if and only if both constraints are held. A unique equilibrium exists when any member is irreplaceable by a larger coalition. It means that, if all nonsignatories would like to replace the player $n^*$ with the smallest marginal benefit of abatement $(\gamma_{n^*})$ in the coalition, the coalition would collapse. In other words, a $(N-1)$-member without player $n^*$ is unprofitable. We can write it in an inequality

$$1 > \sum_{i=1}^{n^*-1} \gamma_i + \sum_{j=n^*+1}^{N} \gamma_j$$

To add the marginal benefit of abatement of player $n^*$ in both sides, the unique equilibrium condition is rewritten as

$$1 + \gamma_{n^*} > \sum_{i=1}^{N} \gamma_i$$

■

## A. Appendix 2

**Proof.** The utility of a signatory $i$ with a $n$-member coalition can be extended to the function with the degree of inequality-aversion as

$$u_i^s(n) = \pi_i^s(n) - \frac{\alpha_i}{N-1}\sum_{m\neq i}\max\left[\pi_m - \pi_i^s(n), 0\right] \tag{14}$$

Because of the external constraint, any nonsignatory has higher utility than what a signatory has. Signatory $i$ has the disadvantage term but no advantage term.

On the other hand, the welfare function of a nonsignatory $j$ with $n$-member coalition is

$$u_j^{ns}(n) = \pi_j^{ns}(n) - \frac{\alpha_j}{N-1}\sum_{j\neq m}\max\left[\pi_m - \pi_j^{ns}(n), 0\right] - \frac{\beta_j}{N-1}\sum_{m\neq j}\max\left[\pi_j^{ns}(n) - \pi_m, 0\right] \tag{15}$$

Nonsignatories could have both the advantage and disadvantage terms. They are advantaged since their individual payoffs are definitely higher than that of a signatory. The one with the highest marginal benefit of the total abatement yields the highest payoff among others. Any other nonsignatory would be disadvantaged to this country.

The stability of the coalition formation depends on the internal and the external constraints. The internal one can be displayed as

$$u_i^s(n^*) > u_i^{ns}(n^*-1)$$
$$\Longrightarrow \left(-1 + \sum_{i=1}^{n^*}\gamma_i\right) - \frac{\alpha_i}{N-1}\sum_{j=n^*+1}^{N}\left[n^*\gamma_j - \left(-1 + \sum_{i=1}^{n^*}\gamma_i\right)\right] > 0$$

The left-hand-side is the utility when $i$ joins the coalition, and the right-hand-side is the utility when $i$ does not join.

If $i$ is not strong inequality averse, the player would follow the internal constraint and decide to participate the coalition. If $i$ is strong inequality averse, both the individual inequality-averse factor $\alpha_i$ and the disadvantage loss are high enough, the player would violate the internal constraint and the consequence is a collapse coalition.

On the other hand, the external constraint can be extended as

$$
\begin{aligned}
& u_j^{ns}\left(n^*\right) > u_i^s\left(n^*+1\right) \\
\Longrightarrow \quad & n^*\gamma_k - \frac{\alpha_k}{N-1}\sum_{k\neq j}\max\left[\pi_j - \pi_k^{ns}\left(n^*\right), 0\right] - \frac{\beta_k}{N-1}\sum_{k\neq j}\max\left[\pi_k^{ns}\left(n^*\right) - \pi_j, 0\right] \\
> \quad & \left(-1+\sum_{i=1}^{n^*}\gamma_i + \gamma_k\right) - \frac{\alpha_k}{N-1}\sum_{k\neq j}\max\left[\left(n^*+1\right)\gamma_j - \left(-1+\sum_{i=1}^{n^*}\gamma_i + \gamma_k\right)\right]
\end{aligned}
$$

where $k$ is a player belongs to $[n^*+1, N]$. The left-hand-side is $k$'s utility when $k$ is a nonsignatory and have the disadvantage loss from higher marginal benefit nonsignatories as well as the advantage loss from all signatories and lower marginal benefit nonsignatories. The right-hand-side is $k$'s utility when $k$ is a signatory which only has the disadvantage loss.

When $k$ does not have enough advantage averse, the player would follow the external constraint and not to participate in the coalition. When $k$ has strong inequality aversion, both the individual inequality-averse factor $\alpha_k$ and $\beta_k$, and the disadvantage and advantage loss are high, the player would violate the external constraint and join the coalition.

To summarise, given all subjects' inequality aversion is not strong enough, both the internal and external constraint are held. There exists a unique stable $n^*$-member coalition as we yield in Proposition 2. If the internal constraint is held, but the external constraint is violated, there exists a stable coalition which the size is larger than $n^*$ members. If the internal constraint is violated, due to any subject having strong inequality aversion, there exists no coalition to be formed.

| Round | Option 1 | Option 2 |
|---|---|---|
| 1 | (£2.5, £2.5) for sure | (£0, £5) with probability 0% and (£5, £0) with probability 100% |
| 2 | (£2.5, £2.5) for sure | (£0, £5) with probability 10% and (£5, £0) with probability 90% |
| 3 | (£2.5, £2.5) for sure | (£0, £5) with probability 20% and (£5, £0) with probability 80% |
| 4 | (£2.5, £2.5) for sure | (£0, £5) with probability 30% and (£5, £0) with probability 70% |
| 5 | (£2.5, £2.5) for sure | (£0, £5) with probability 40% and (£5, £0) with probability 60% |
| 6 | (£2.5, £2.5) for sure | (£0, £5) with probability 50% and (£5, £0) with probability 50% |
| 7 | (£2.5, £2.5) for sure | (£0, £5) with probability 60% and (£5, £0) with probability 40% |
| 8 | (£2.5, £2.5) for sure | (£0, £5) with probability 70% and (£5, £0) with probability 30% |
| 9 | (£2.5, £2.5) for sure | (£0, £5) with probability 80% and (£5, £0) with probability 20% |
| 10 | (£2.5, £2.5) for sure | (£0, £5) with probability 90% and (£5, £0) with probability 10% |
| 11 | (£2.5, £2.5) for sure | (£0, £5) with probability 100% and (£5, £0) with probability 0% |

Table 1: Distribution of payoff in all 11 rounds in the inequality-aversion test

■

| Round | Player 1 | Player 2 | Player 3 | Player 4 | Player 5 |
|---|---|---|---|---|---|
| $1-15$ | $0.675^*$ | $0.375^*$ | 0.125 | 0.10 | 0.075 |
| $16-30$ | 0.075 | $0.15^*$ | $0.25^*$ | $0.3^*$ | $0.35^*$ |
| $31-45$ | $0.40^*$ | $0.65^*$ | 0.075 | 0.10 | 0.125 |
| $46-60$ | 0.05 | 0.1 | $0.4^*$ | $0.35^*$ | $0.3^*$ |
| * means the weakly dominant strategy of the player is joining the coalition. | | | | | |

Table 2: List of parameters of marginal benefit for players taking Treatment 1

| Round | Player 1 | Player 2 | Player 3 | Player 4 | Player 5 |
|---|---|---|---|---|---|
| $1-15$ | 0.075 | 0.1 | $0.45^*$ | $0.35^*$ | $0.25^*$ |
| $16-30$ | 0.125 | 0.1 | 0.15 | $0.5^*$ | $0.55^*$ |
| $31-45$ | $0.45^*$ | $0.6^*$ | 0.05 | 0.2 | 0.1 |
| $46-60$ | $0.45^*$ | $0.25^*$ | $0.2^*$ | $0.15^*$ | 0.05 |
| * means the weakly dominant strategy of the player is joining the coalition. | | | | | |

Table 3: List of parameters of marginal benefit for players taking Treatment 2

| Variable | Inequality-aversion level<br>OLS Regression | | |
|---|---|---|---|
| Constant term | $-12.53$<br>(11.15) | | |
| AGE | 0.007<br>(0.006) | | |
| POLITIC | 0.005<br>(0.03) | | |
| RELIGION | $-0.02$<br>(0.02) | | |
| Log Likelihood | 19.13514 | R-squared | 0.042 |
| Total Observation | 50 | | |
| Note: Each cell contains coefficient and standard error in parenthesis. | | | |
| *, **, *** are significant at 10%, 5%, and 1% respectively. | | | |

Table 4: OLS estimation of inequality-averse preference

| Variable | Probit MLEs(1) | Probit MLEs(2) | Probit MLEs(3) | Probit MLEs(4) | Probit MLEs(5) |
|---|---|---|---|---|---|
| Constant term | 8.32 | 0.52*** | -9.77 | -0.05 | 11.01 |
| | (12.49) | (0.16) | (20.54) | (0.05) | (16.72) |
| DECISION (-1) | 1.19*** | | 1.36*** | | 1.01*** |
| | (0.07) | | (0.13) | | (0.09) |
| INEQ | 0.50*** | 0.81*** | | -0.15** | |
| | (0.19) | (0.24) | | (0.08) | |
| AGE | -0.005 | | 0.005 | | -0.005 |
| | (0.006) | | (0.01) | | (0.008) |
| POLITIC | 0.05 | | -0.13** | | 0.23*** |
| | (0.03) | | (0.05) | | (0.05) |
| RELIGION | -0.05** | | 0.02 | | -0.17*** |
| | (0.02) | | (0.03) | | (0.03) |
| WD STRATEGY | 1.16*** | | | | |
| | (0.10) | | | | |
| $\gamma$ | -1.27*** | | | | -6.45*** |
| | (0.26) | | | | (1.11) |
| TC (-1) | -0.16 | | -0.26 | | -0.36** |
| | (0.12) | | (0.21) | | (0.16) |
| Log Likelihood | -1165.01 | -621.21 | -515.43 | -769.35 | -629.48 |
| Total Observation | 2520 | 1500 | 1400 | 1120 | 1120 |
| Observation with decision is 'Join' | 1692 | 1279 | 1185 | 507 | 507 |

Note: Each cell contains coefficient and standard error in parenthesis.
*, **, *** are significant at 10%, 5%, and 1% respectively.

Table 5: Probit estimations of probability of joining a coalition

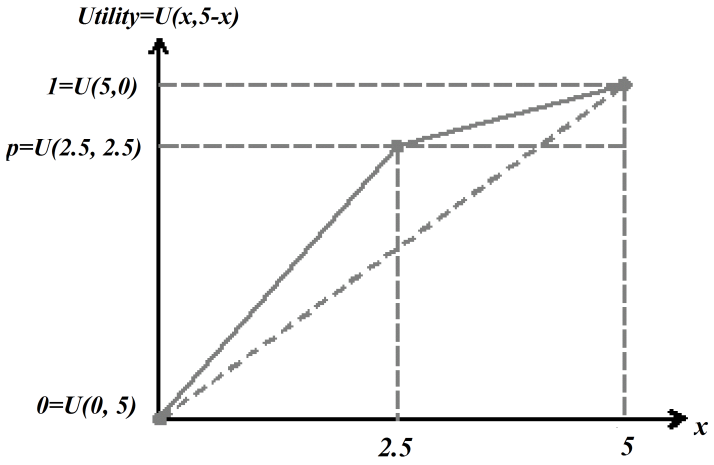


Figure 1: Subject A's inequality-averse preference

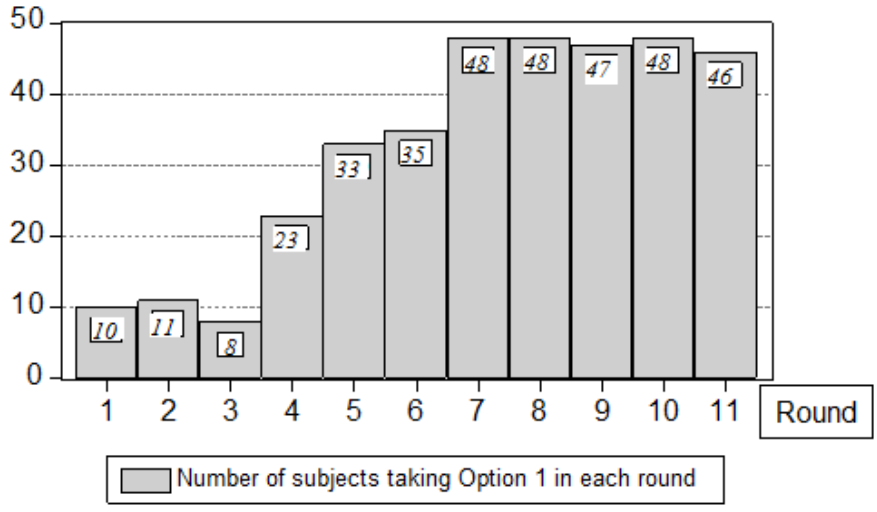


Figure 2: Number of subjects taking 'Option 1' in each round

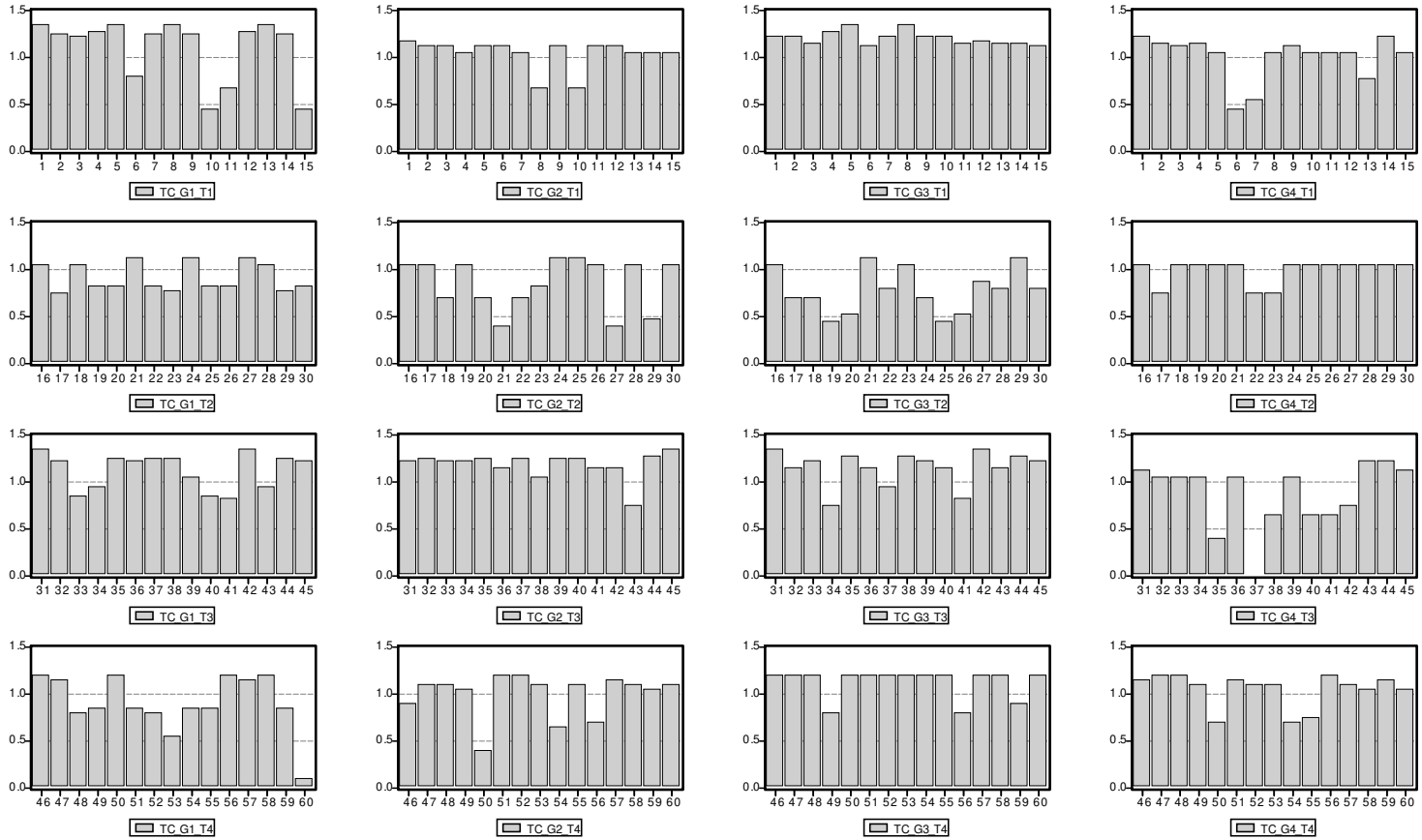


Figure 3: The total contribution of Group 1-4 in four sub-treatments

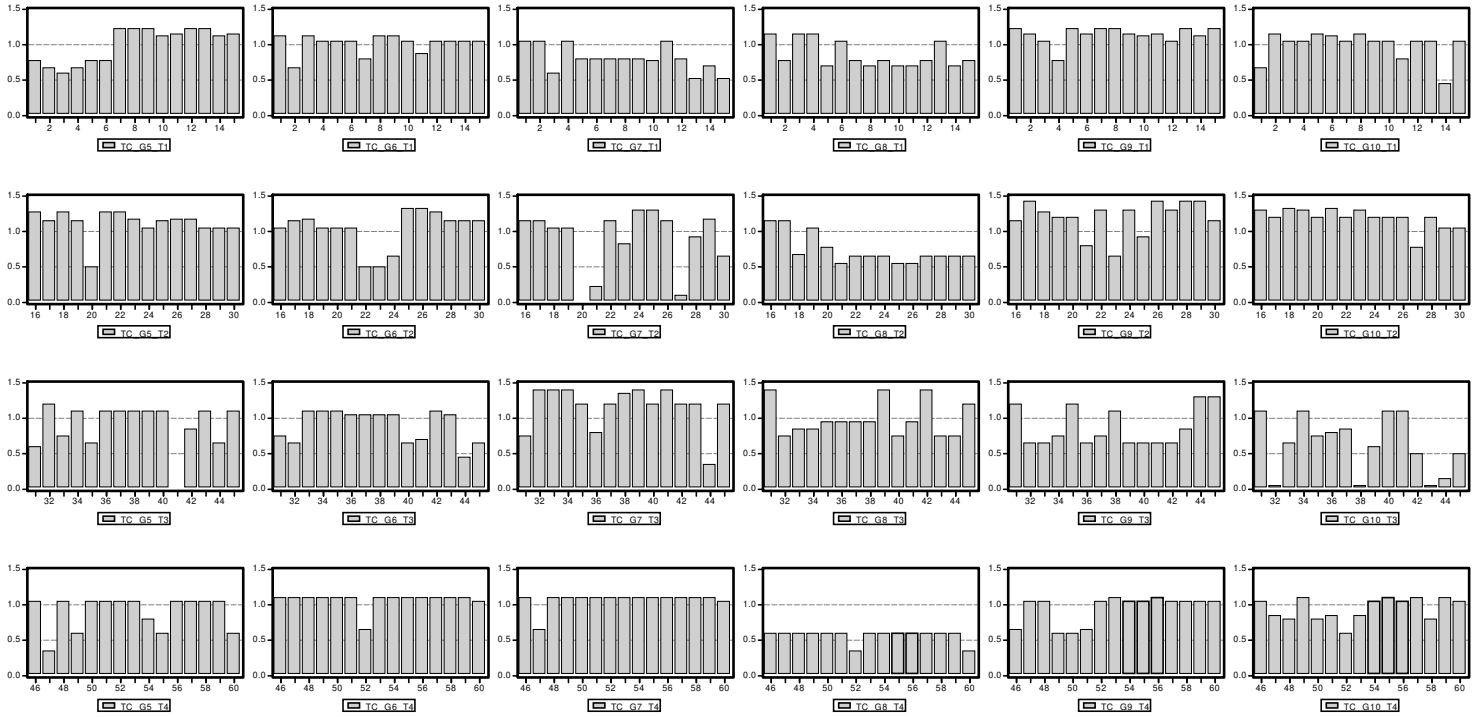


Figure 4: The total contribution of Group 5-10 in four sub-treatments

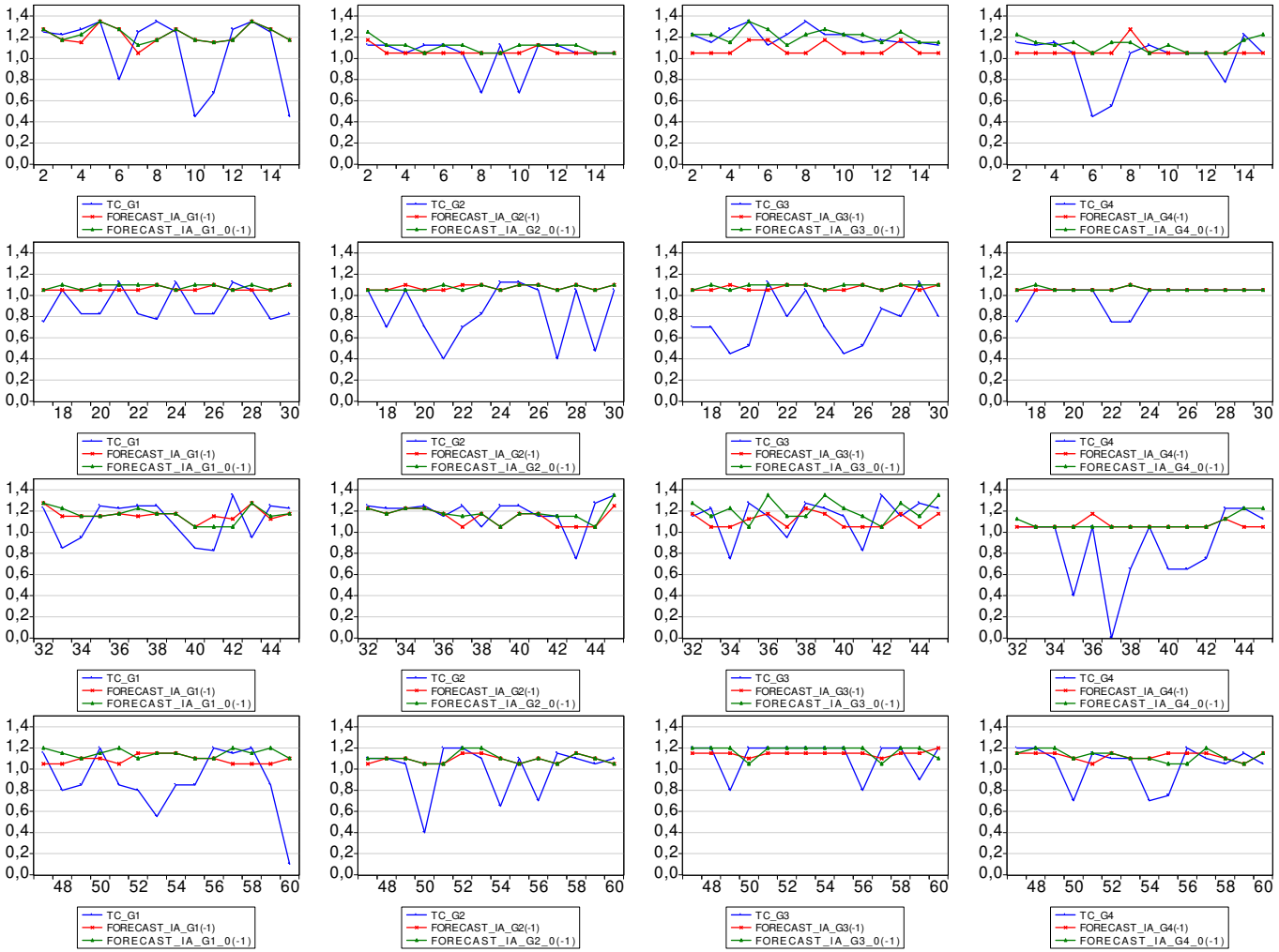


Figure 5: The actual total contribution and the predicted total contribution of Groups 1 to 4

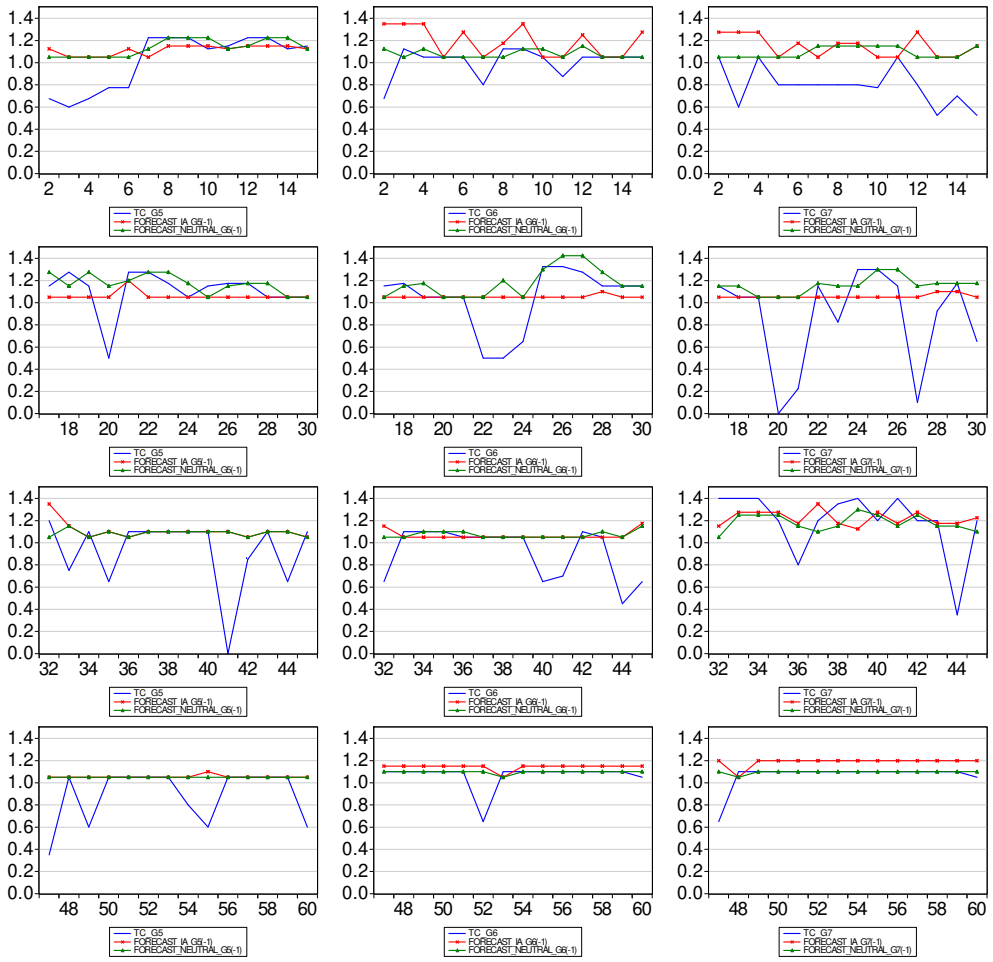


Figure 6: The actual total contribution and the predicted total contribution of Groups 5 to 7

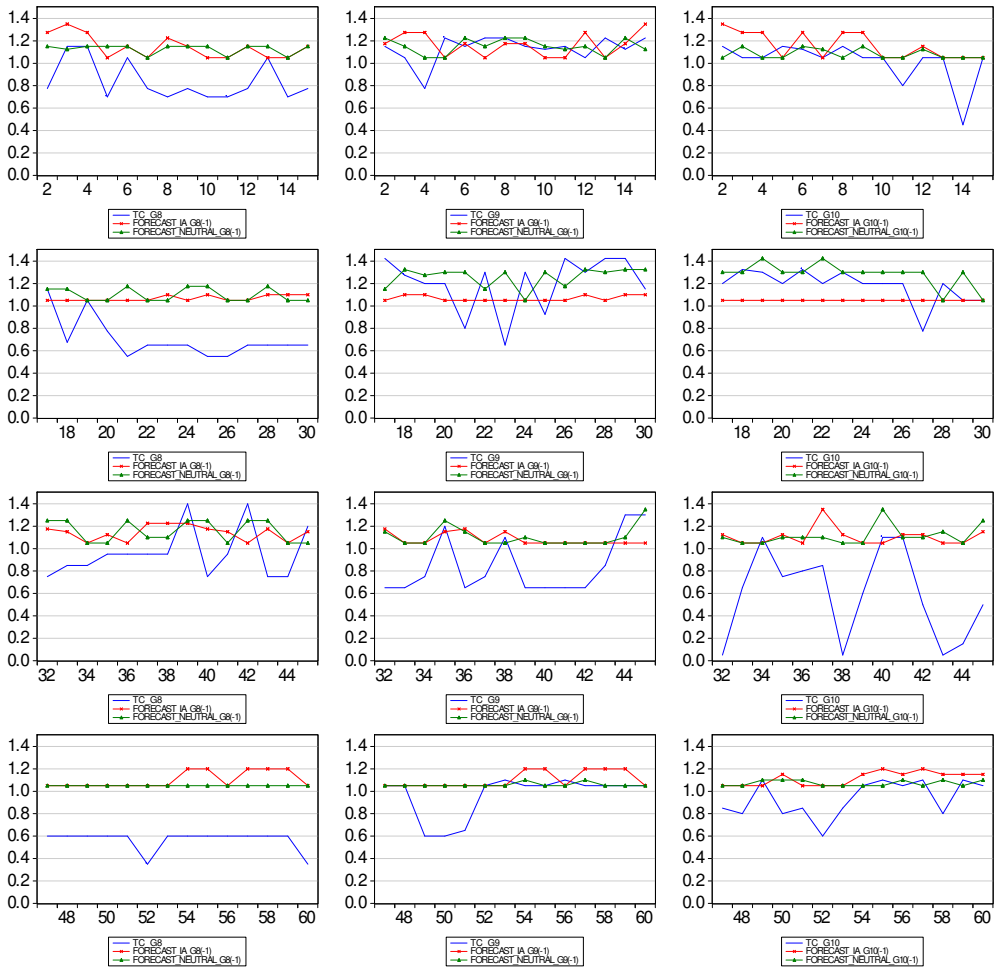


Figure 7: The actual total contribution and the predicted total contribution of Groups 8 to 10

## References


[1] Barrett, S. (1994). "Self-Enforcing International Environmental Agreements." *Oxford Economic Paper*. Vol. 46(1): 878-894.

[2] Barrett, S. (2001). "International Cooperation for Sale." *European Economic Review*. Vol. 45: 1835-1850.

[3] Blanco, M., Engelmann, D. and Normann, H.T. (2011). "A Within-Subject Analysis of Other-Regarding Preferences". *Games and Economic Behavior*. Vol. 72(2): 321-338.

[4] Breton, M.; Sbragia, L.; and Zaccour, G. (2010). "A Dynamic Model for International Environmental Agreements". *Environmental and Resource Economics*. Vol. 45(1): 25-48.

[5] Burger, N.E. and Kolstad, C.D. (2009). "Voluntary Public Goods Provision, Coalition Formation, and Uncertainty" NBER Working Papers 15543, National Bureau of Economic Research, Inc.

[6] Carlsson, F.; Daruvala, D. and Johansson-Stenman, O. (2005). "Are People Inequality-Averse, or Just Risk-Averse?" *Economica*. Vol. 72(3): 375-396.

[7] Charness, G. and Rabin, M. (2002). "Understanding Social Preferences With Simple Tests". *The Quarterly Journal of Economics*. Vol. 117(3): 817-869.

[8] Dannenberg, A., Riechmann, T., Sturm, B. and Vogt, C. (2007). "Inequity Aversion and Individual Behavior in Public Good Games: An Experimental Investigation". Working paper.

[9] D'Aspremont, C., Jacquemin, A., Gabszewicz, J., Weymark, J. (1983). "On the Stability of Collusive Price Leadership". *Canadian Journal of Economics*. Vol. 16(1): 17-25.

[10] Fehr, E. and Schmidt, K. M. (1999). "A Theory Of Fairness, Competition, And Cooperation". *The Quarterly Journal of Economics*. Vol. 114(3): 817-868.

[11] Fischbacher, U. (2007). "z-Tree: Zutich Toolbox for Ready-made Economic Experiments." *Experimental Economics*. Vol. 10(2): 171-178.

[12] Greiner, B. (2004). "The Online Recruitment System ORSEE 2.0 - A Guide for the Organization of Experiments in Economics." Working Paper Series in Economics 10, University of Cologne, Department of Economics.

[13] Kolstad, C. D. (2014) "Public Goods Agreements with Other-Regarding Preferences". Working paper

[14] Kroll, Y. and Davidovitz, L. (2003). "Inequality Aversion versus Risk Aversion". *Economica*. Vol. 70 (277): 19-29.

[15] Kosfeld, M., Okada, A. and Riedl, A. (2009). "Institution Formation in Public Goods Games." *American Economic Review*. Vol. 99(4): 1335–55.

[16] Yang, Y.; Onderstal, S.; and Schram, A. (2012). "Inequity Aversion Revisited". Working paper.