Print ISSN: 2671-4981 / Online ISSN: 2671-499X
doi:10.13106/jbees.2020.vol10.no2.5

The Influence of Reciprocity on Individual Decisions in a Climate Coalition Experiment*

Yu-Hsuan LIN**

Received: October 16, 2019. Revised: December 01, 2019. Accepted: April 05, 2020.

Abstract

Purpose: This study examines the impact of individual reciprocal preferences on coalition formation. The reciprocal model considers a player's own payoff, the player's perception of others' payoffs, and others' perceptions of the player's payoff. **Research design, data and methodology:** A reciprocal model is built to illustrate how reciprocity influences individual decisions in a coalition game and its formation. The prediction is examined with experimental evidences from a dictator game and a membership game. **Results:** The theoretical result suggests that the coalition formation could be unstable due to negative reciprocal kindness. The experimental findings support that negative reciprocal kindness could lead players participating in a coalition, no matter their dominant strategies are. When subjects were essential to make contributions to a coalition, they were more likely to cooperate if they were treated badly. In contrast, when subjects were unnecessary, the reciprocal kindness could enhance cooperative tendencies. **Conclusions:** This study reveals that the reciprocal behavior could influence individual decisions and reshape the coalition formation. In terms of policy implications, this study has shown that coalition formation could be reshaped by reciprocal prefe rences. Due to the strategic and complicated decision process in an interactive environment, a comprehensive investigation of factors would be required in a climate coalition in practice.

Keywords : Reciprocity, Social Preference, Climate Coalition, International Environmental Agreement, Experimental Economics

JEL Classification Code: C91, D64, H41, Q54

1. Introduction

The dynamics of how international environmental agreements are reached have been discussed for decades, with an important subset of such research focusing on climate coalitions in particular (e.g., Carraro, 1999; Carraro, Eyckmans, & Finus, 2006; Nagashima, Dellink, Van Ierland, & Weikard, 2009). A seminal study by Barrett (1994) hypothesized that because countries are self-interested, their participation in such coalitions will tend to be self-enforced. Most of the subsequent literature, notably including Breton, Sbragia, and Zaccour (2010), Bosetti, Carraro, De Cian, Massetti, and Tavoni (2013), and Nordhaus (2015) has suggested that stable coalitions achieve little, if their agreements include no additional policy mechanisms.

However, a number of experimental studies have challenged this thinking, by suggesting that cooperation does exist in the absence of policy interventions (such as Kosfeld, Okada, & Riedl, 2009; Bosetti, Heugues, & Tavoni, 2017; Calzolari, Casari, & Ghidoni, 2018). These studies have claimed that people are more likely to cooperate than the self-interested prediction suggests. Several studies, including those by Charness and Rabin (2002); Fischbacher and Gachter (2010); Hadjiyiannis, İriş, and Tabakis (2012) and Dannenberg, Löschel, Paolacci, Reif, and Tavoni (2015), have reported that *social* or *other-regarding* preferences are the main reason for the formation, in reality, of larger coalitions than theories have predicted.

Several models of social preferences have been proposed. For example, Hahn and Ritz (2014) and Lin (2018) took account of pure altruism, that people may look after not only their own wellbeing but also the wellbeing of others. Lange (2006) and Lin (2017) explored how equity considerations affect countries' cooperation on a climate

*This study was supported by the Catholic University of Korea Research Fund (2020).
**Assistant Professor, Department of Economics, the Catholic University of Korea. 43, Jibong ro, Bucheon, 14662, Republic of Korea, Email: yuhsuan.lin@catholic.ac.kr

coalition formation. These authors assumed that such attitude for others is unidirectional and does not ask for anything in return. Yet, psychological evidence indicates that most altruistic behavior is more complex (Nyborg, 2018; Rabin, 1993): people make decisions based on how they are treated by others, behaving generously when they meet altruistic people and ungenerously when they meet stingy ones.

The motivations that underlie reciprocal behavior have also been studied extensively. Such research can be categorized into three strands: reciprocal fairness (e.g., Fehr & Schmidt, 1999; Bolton & Ockenfels 2000), reciprocal altruism (e.g., Levine, 1998; Dufwenberg & Kirchsteiger, 2004), and the quest for efficiency gains through reciprocity (Brandts & Schram, 2001) - for an overview of all three strands, see (Seinen & Schram, 2006).

The reciprocal model has been tested experimentally, for instance, via ultimatum bargaining games (Dickinson, 2000) and public-goods games (Bardsley & Moffatt, 2007). One of the present paper's distinctive contributions is that it serves as a bridge between such experimental studies and the existing body of literature on the motivations underlying reciprocal altruism. Specifically, we will use an experimental approach to test the prior theoretical literature's consensus that a coalition may be stabilized through sufficiently strong and widespread reciprocal preferences – an idea that can intuitively be questioned as unworkable in practice. The findings provided policy implications on international conventions, climate coalitions in particular. A number of experimental studies, such as Lin (2018) and İriş, Lee, and Tavoni (2019), employed laboratory evidences to explain that climate coalitions are driven more by public pressure than by self-interest. The microfoundation approach is a reasonable tool to explain the behavior of nations.

Previously, Lin (2018) considered the impact of unidirectional altruistic preferences on individuals' decision-making in a climate coalition, using a design of self-interested dominant-strategy equilibrium to inspect both the individual behavior and coalition formation. However, the experimental results showed that the same players who were altruistic in a dictator game became hostile in an interactive climate-coalition game, implying that the unidirectional model may be unable to predict decisions in an interactive game. Thus, to ensure that its model is capable of understanding individuals' cooperative behavior, this study also takes account of mutual social preferences. Specifically, it studies how mutual social preferences influence individual decisions on climate-coalition membership, based on experimental evidence, as well as seeking to explain how individual social preferences affect coalition formation.

This study examines the impact of individual reciprocal preferences on coalition formation. The reciprocal model considers a player's own payoff, the player's perception of others' payoffs, and others' perceptions of the player's payoff. With the addition of reciprocal preferences, in theory, coalition formation can be reshaped and moved beyond the dominant-strategy equilibrium. Specifically, due to the interactive nature of perceptions of players' payoffs in this scenario, coalition size might become either smaller or larger than they would in a state of dominant-strategy equilibrium based on self-interest. That is, people learn about their feelings based on their respective histories of prior interaction and will make decisions, in part, based on how they have been treated by others. We hypothesize that negative reciprocal kindness could turn down players' dominant strategies, no matter whether they are or not critical to an effective coalition.

This study also employs Lin (2018) experimental evidence to test the theoretic prediction. The experiment consists of a dictator game and a membership game. The later one concentrates on a dominant strategy equilibrium. Its key strength is that it allows investigation of individual incentives for participating in a coalition. Our findings suggest that negative reciprocal kindness would lead players participating in a coalition, no matter whether they were critical to an effective coalition or not.

The remainder of this paper is structured as follows: Section 2 describes the reciprocity model and Section 3, the experimental design. Section 4 illustrates reciprocal behavior in both the dictator game and the coalition membership game with experimental evidence, and the final section presents our conclusions

2. Models

2.1. Self-interested model

The model in this study is built for illustrating individual behavior in a climate coalition. In a climate-coalition game played by N countries, the model considers two scenarios: the self-interested and the reciprocal. In the first scenario, countries concern themselves with only their own payoffs. The welfare function of an arbitrary country k is its own payoff matrix for strategy profile ($S_1 \times \dots S_k \dots \times S_N \rightarrow \mathbb{R}$) as

$$u_k\left(s_k, s_{-k}\right) = \pi_k\left(s_k, s_{-k}\right) \quad \forall k \in [1, N] \tag{1}$$

where s_k, a strategy from country k's strategy set S_k, defines country k's decision in participating in a coalition. s_{-k} is used to denote the strategies of countries other than

country k. The strategy profile is set as the dummy variable which equals 1 if the country chooses to join or 0 if the country chooses not to join. Within a curly bracket is a strategy profile which summarizes the collection of all countries' strategies. The self-interested welfare function equals the payoff, which depends on a strategy profile.

Since the purpose of a climate coalition is dedicated to action against climate change, it is intuitive to assume that countries which participate in the coalition (hereafter, '*signatory*') do abatement, while the rest which does not participate (hereafter, '*nonsignatory*') does pollution. Suppose that n countries participate in the coalition, signatories move as one to maximize the coalition's welfare. When no effective coalition is formed, all players receive nothing. An effective coalition is formed when the collective contribution is large enough so the summation of the marginal benefits of signatories is no less than the standard abatement cost. A signatory i chooses to join a coalition ($s_i = 1$) and s_{-i} is used to denote the strategies of countries other than country i. Each signatory shares the joint payoff equally, so that signatory i's payoff is the sum of all signatories' marginal benefit minus the standard abatement cost,

$$\pi_i(s_i = 1, s_{-i}) = \begin{cases} (\sum_{i=1}^{n} \gamma_i) - 1 & \text{if } \sum_{i=1}^{n} \gamma_i \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in [1, n] \tag{2}$$

where γ_i is signatory i's marginal benefit of total abatement in the range of 0 and 1.

This implies that a country has incentive to abate only if the overall signatories' abatement benefit is large enough to overcome the private abatement cost

On the other hand, a nonsignatory j chooses not to join a coalition ($s_j = 0$) and s_{-j} is used to denote the strategies of countries other than country j. Nonsignatory j pursues its own interest by performing no abatement. When an effective coalition is formed, the payoff of nonsignatory j is the product of its individual benefit and the coalition size,

$$\pi_j(s_j = 0, s_{-j}) = \begin{cases} \gamma_j n & \text{if } \sum_{i=1}^{n} \gamma_i \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad \forall j \in [n+1, N] \tag{3}$$

where γ_j is nonsignatory j's marginal benefit of total abatement in the range of 0 and 1.

Following d'Aspremont, Jacquemin, Gabszewicz, and Weymark (1983), if all countries are self-interested, a stable n-member coalition exists when the following constraints are satisfied:

$$u_i(s_i = 1, s_{-i}) \geq u_i(s_i = 0, s_{-i}) \tag{4}$$

$$u_j(s_j = 0, s_{-j}) \geq u_j(s_j = 1, s_{-j}) \tag{5}$$

Inequality (4) is the internal constraint which ensures a signatory has no incentive to leave an effective coalition and becomes a non-signatory. On the other hand, inequality (5) is the external constraint which ensures that any nonsignatory has no incentives to join the coalition as its new member. Thus, when both constraints are satisfied, an n-member stable coalition would exist.

Though the coalition formation is predictable, many stable coalitions have been identified by both theoretical and experimental studies (e.g., Kosfeld et al., 2009). In these cases, the sizes of the coalitions tended to be predictable, but the joining decisions of individual countries were not. However, this was probably due to a lack of clear-cut preference on the part of the countries, which made their decisions difficult or impossible to foresee. To help rectify that problem, a special property of stable coalitions is given in the following result.

Proposition 1: Consider the self-interested behavior of a coalition game, a dominant strategy equilibrium is the only stable coalition.

Proof: Since countries are self-interested, the welfare function is the country's own payoff function. By setting up the ranks of marginal benefits $1 > \gamma_1 > \cdots > \gamma_{n^*-1} > \gamma_{n^*} > \gamma_{n^*+1} > \cdots \gamma_N > 0$, a coalition is stable when both internal and external constraints are satisfied. Thus, the constraints (4) and (5) could be rewritten in the payoff function as

$$\left(\sum_{i=1}^{n^*} \gamma_i\right) - 1 \geq 0 \qquad \forall i \in [1, n^*] \tag{4'}$$

$$\gamma_j n^* \geq \left(\sum_{i=1}^{n^*} \gamma_i\right) + \gamma_j - 1 \quad \forall j \in [n^*+1, N] \tag{5'}$$

The left-hand-side of inequality (4') is a signatory's payoff whilst the right-hand-side is the payoff when it becomes a nonsignatory. The left-hand-side of inequality (5') is the payoff of a nonsignatory while the right-hand-side is its payoff when it becomes a signatory. Inequality (5') also implies that a nonsignatory has a higher payoff than what a signatory has. In other words, the free-riding benefit could ensure the stability externally.

A dominant strategy equilibrium categorizes countries into two groups: *critical* (country $1, \ldots, n^*$) and *non-critical* (country $n^*+1, \ldots, N$). No matter what strategy is chosen by others, the critical countries will choose to cooperate

because they are necessary to the existence of an effective coalition, and non-critical countries will choose not to participate because they could receive free-riding advantages. A dominant strategy equilibrium implies that a critical country cannot be replaced, even by all of the non-critical countries combined. So that

$$\gamma_{n^*} > \sum_{j=n^*+1}^{N} \gamma_j \quad (6)$$

In summary, a dominant strategy equilibrium ensures a n^*-member stable coalition in the game. Those countries with large marginal benefits are essential to an effective coalition and have no incentive to separate. Those countries with small marginal benefits are not necessary and have free-riding incentives. Thus, this profile is the only stable formation.

2.2. Reciprocal model

Turning now to consider a reciprocity model, different from the welfare function (1), the reciprocal welfare depends on not only the combinations of actual strategies but also decision makers' beliefs. Following Rabin (1993), we assume that country k's subjective expected welfare when it chooses its strategy s_k depends on three factors: (i) country k's actual strategy, (ii) its beliefs about the other countries' strategy choice, and (iii) its beliefs about the other countries' beliefs about its strategy. Thus, we will use the following notations: $a_k \in S_k$ is denoted as the actual strategies chosen by country k; $b_{-k} \in S_{-k}$ is denoted as country k's belief about which strategy other countries are choosing; and $c_k \in S_k$ is denoted as country k's belief about what other countries believe its own strategy is.

Country k attempts to maximise its expected welfare, which incorporates both payoff and its shared notion of fairness. This reciprocal welfare can be expressed as

$$\begin{aligned} & v_k(a_k, b_{-k}, c_k) \\ &= v_k^s(a_k, b_{-k}) + v_{-k}^r(c_k, b_{-k})\left[1 + v_k^a(a_k, b_{-k})\right] \\ & \qquad \forall k \in [1, N] \end{aligned} \quad (7)$$

where

v_k^s is the payoff which depends on k's actual strategy and k's beliefs about what strategies others are choosing,

v_k^r is reciprocal kindness which depends on k's beliefs about what others believe k's strategy is and k's beliefs about what strategies others are choosing,

v_k^a is straight kindness which also depends on k's actual strategy and k's beliefs about what strategies others are choosing.

Both kindness functions are normalized, straight kindness and reciprocal kindness occur within a range from -1 to 0.5.

In the same way as the welfare function (1), country k's pure self-interest (v_k^s) is country k's own payoff. The notion of fairness is used to specify country k's preference by both reciprocal kindness (v_k^r) and straight kindness (v_k^a). Reciprocal kindness indicates how country k experiences other players' kindness, while straight kindness indicates its kindness to other players. The impact of reciprocal kindness on welfare depends on a country's feelings about others. That is, the overall welfare of who feel they are being treated badly will be lower than those countries' payoffs. Straight kindness, on the other hand, is rooted in the strength of feeling. If a country is straight hostile, it cares little about others' decisions, while a country who is straight generous depends on the communication of kindness for its welfare – i.e., if treated kindly, its welfare is higher than its payoff, and if treated badly, it is lower. These definitions of kindness are expressed as follows.

Definition 1: Reciprocal kindness defines country k's beliefs about how generous other countries are being to it, as

$$v_{-k}^r(c_k, b_{-k}) \equiv \frac{\pi_k(c_k, b_{-k}) - \pi_k^e(c_k)}{\pi_k^h(c_k) - \pi_k^{min}(c_k)} \quad (8)$$

if $\pi_k^h(c_k) - \pi_k^{min}(c_k) = 0$, then $u_k^r(c_k, b_{-k}) = 0$.

Country k's reciprocal kindness consists of its payoffs: $\pi_k(c_k, b_{-k})$ is the payoff that k thinks what others believe in k's strategy and others think what k believes in their strategies. $\pi_k^h(c_k)$ and $\pi_k^l(c_k)$ are the highest and lowest in the set of all feasible Pareto-efficient payoffs, respectively. $\pi_k^e(c_{-k})$ is the *equitable payoff*, defined as the average of the highest and lowest payoffs, $[\pi_k^h(c_k) + \pi_k^l(c_k)]/2$. The equitable payoff provides a crude reference point to measure how kind others are being to country k. Finally, $\pi_k^{min}(c_k)$ is the worst possible payoff for player k in the set of all possible payoffs.

Definition 2: Straight kindness defines player k's kindness to other non-k countries, as

$$v_k^a(a_k, b_{-k}) \equiv \frac{\pi_{-k}(a_k, b_{-k}) - \pi_{-k}^e(b_{-k})}{\pi_{-k}^h(b_{-k'}) - \pi_{-k}^{min}(b_{-k})} \quad (9)$$

if $\pi_{-k}^h(b_{-k},) - \pi_{-k}^{min}(b_{-k}) = 0$, then $u_k^a(a_k, b_{-k}) = 0$.

In other words, country k's straight kindness consists of the payoffs of other countries: i.e., $\pi_{-k}(b_{-k}, a_k)$ is a non-k country's payoff that what strategy k chooses and

what k believes others would do to her; $\pi_{-k}^{h}(b_{-k},\ a_k)$ and $\pi_{-k}^{l}(b_{-k},)$ are the highest and lowest payoffs, respectively, that a non-k player could receive from among the set of all feasible Pareto-efficient payoffs; the equitable payoff, $\pi_{-k}^{e}(b_{-k})$, is the average of the highest and lowest payoffs; and $\pi_k^{min}(c_k)$ is the worst possible payoff for a non-k player in the set of all possible payoffs.

Using these definitions, the following proposition is proposed to discuss the effects of reciprocal behavior on coalition formation.

Proposition 2: Considering the reciprocal behavior in a coalition game, a dominant strategy equilibrium may become unstable internally.

Proof. Having defined the reciprocal preference, we now find some examples which may change the coalition formation internally and externally. When we consider reciprocal behavior in the welfare function (7), the function depends on countries' decisions and beliefs. A strategy of country k (S_k) is to determine whether or not to participate in a coalition. A dummy variable denotes the status of strategies ($S_k = 1$ means k chooses to cooperate and $S_k = 0$ means k chooses not to cooperate).

Considering a situation in which an effective coalition is formed, the highest in the set of all feasible Pareto-efficient payoffs of a critical country i is $(\sum_{i=1}^{N}\gamma_i - 1)$ when all countries participate in a coalition. By contrast, the lowest in the set of all feasible Pareto-efficient payoffs is $(\sum_{i=1}^{n^*}\gamma_i - 1)$ when only critical countries participate. Thus, the equitable payoff is $\frac{\left(\sum_{i=1}^{n^*}\gamma_i+\sum_{i=1}^{N}\gamma_i-2\right)}{2}$. The worst possible payoff is zero, which means no effective coalition is formed.

Meanwhile, the highest in the set of all feasible Pareto-efficient payoffs of a non-critical country j is $(N-1)\gamma_j$ when country j separates. By contrast, the lowest in the set of all feasible Pareto-efficient payoffs is $n^*\gamma_j$ when it joins the smallest effective coalition. Thus, the equitable payoff for a non-critical country is $\frac{(N+n^*-1)\gamma_j}{2}$. The worst possible outcome occurs when no effective coalition is formed, and everyone gets nothing.

Regarding the internal stability, the reciprocal welfare function of a critical country i is

$$v_i(a_i\ , b_{-i}, c_i\ ) = \pi_i\ (a_i\ , b_{-i}) + \frac{\pi_i\ (c_i\ ,b_{-i})-\pi_i^e(c_i\ )}{\pi_i^h(c_i\ )-\pi_i^{min}(c_i\ )}\left[1+\frac{\pi_{-i}(a_i\ ,b_{-i})-\pi_{-i}^e(b_{-i})}{\pi_{-i}^h(b_{-i'})-\pi_{-i}^{min}(b_{-i})}\right] \quad (10)$$

In a situation of a dominant strategy equilibrium, a critical country i believes a non-critical country j does not cooperate and i also believes j would believe that i cooperates. When i actually cooperates, country i's reciprocal welfare is

$$v_i(a_i\ = 1, b_{-i} = 0, c_i\ = 1) = \left(\sum_{i=1}^{n^*}\gamma_i - 1\right) - \frac{\sum_{i=n^*+1}^{N}\gamma_i}{2\left(\sum_{i=1}^{N}\gamma_i-1\right)}\left[1-\frac{(N-n^*-1)}{(N-1)}\right] \quad (11)$$

If country i chooses not to cooperate, the reciprocal welfare becomes

$$v_i(a_i\ = 0, b_{-i} = 0, c_i\ = 1) = -\frac{\sum_{i=n^*+1}^{N}\gamma_i}{2\left(\sum_{i=1}^{N}\gamma_i-1\right)}\left[1-\frac{(N+n^*-1)}{2(N-1)}\right] \quad (12)$$

No matter country i chooses to cooperate or not, both reciprocal and straight kindness are negative. Compare (11) with (12), if the following condition occurs, the internal stability would be violated:

$$4\left(\sum_{i=1}^{N}\gamma_i - 1\right)\left(\sum_{i=1}^{n^*}\gamma_i - 1\right)(N-1) < \sum_{i=n^*+1}^{N}\gamma_i\ (3n^* + 1 - N) \quad (13)$$

It means that if other non-critical countries are unkind to country j, the straight kindness is still not enough no matter i's choice. On the other hand, because the reciprocal kindness is negative, the internal stability could be broken by a critical country. In other words, this critical country may turn down a coalition due to its hostile feeling about unkind non-critical countries.

Having discussed the internal stability, the external stability could be violated by a non-critical country j. Considering an effective coalition, the highest in the set of all feasible Pareto-efficient payoffs of a non-critical country j is $(N-1)\gamma_j$ if country j separates from the grand coalition. By contrast, the lowest in the set of all feasible Pareto-efficient payoffs is $(n^*\gamma_i)$ when it joins the smallest effective coalition. Thus, the equitable payoff for a non-critical country is $\frac{(N+n^*-1)\gamma_j}{2}$. The worst possible outcome occurs when no effective coalition is formed, and everyone gets nothing.

In a situation of a dominant strategy equilibrium, a non-critical country j believes a critical country i does not cooperate and j also believes that i believes j separates. When country j actually separates from an ineffective coalition, j's reciprocal welfare is

$$v_j\ (a_j\ = 0, b_{-j} = 0, c_j\ = 0) = -\frac{(N+n^*-1)}{2(N-1)} \quad (14)$$

If non-critical j believes that no effective coalition would be formed, everyone gets nothing. Country j's reciprocal kindness becomes negative, and its straight kindness is zero. A negative welfare level implies that the non-critical country is unkind to others but feels hostile from other non-critical countries.

If non-critical j actually chooses to participate an ineffective coalition and its reciprocal welfare becomes

$$v_j\ \left(a_i\ = 1, b_{-i} = 0, c_i\ = 0\right) = -\frac{(N+n^*-1)}{2(N-1)} \qquad (15)$$

Country j receives a negative welfare level due to its beliefs about the separation of a critical country. Comparing (14) with (15), no matter what country j's actual strategy is, it has identical reciprocal welfare if j behaves unkind and feels hostile. The external stability, therefore, becomes unstable.

In summary, taking the reciprocal behavior into account, a stable coalition could be reshaped internally and externally. The internal stability could be broken by the hostile feeling about others, and the external stability could be violated due to the coalitional benefit.

3. Experimental Design

This study employs the laboratory experiment results reported by Lin (2018) to test the hypotheses from the reciprocal model. Here we briefly introduce the design of the experiment. Fifty subjects were recruited at a laboratory in a University in the North East of England. That experiment consisted of two parts: the first being a dictator game that evaluated individual altruistic attitudes, and the second, a membership game mimics a climate coalition formation. The design in other experimental studies may be able to observe the possibility of multiple equilibria; however, they were incapable to predict individual decisions in an interactive game. Therefore, the design of dominant-strategy equilibrium provides a suitable environment in which to observe individual decisions, because it ensures that each player's assigned strategy provides a better payoff than any other strategy regardless of the other players' strategy.

Following up on the previous section, two hypotheses are, therefore, proposed:

Hypothesis 1: Negative reciprocal kindness and positive straight kindness could lead a critical country staying away from a coalition.

Hypothesis 2: Negative reciprocal kindness and positive straight kindness could lead a non-critical country participating in a coalition.

In the dictator game, subjects were anonymously and randomly paired with each other to make 20 'keep-or-give' decisions. In each round, each subject was given one token, and required to decide whether to give it to her/his partner. The participants did not learn what their partners' decisions were until the end of the session. Each of the 20 rounds featured different monetary values for keeping the token and giving it away along with how many subjects decided to give it away in each case. Because the subjects in the dictator game did not know how they were being treated by their partners, only straight kindness – not reciprocal kindness – was calculated. And because the decisions to keep and to give are both Pareto-efficient solutions, the worst payoff for the opponent was nothing. Thus, a subject's straight kindness level was indicated as either -0.5 (keep) or 0.5 (give) in the dictator game.

Turning now to the membership game, subjects were randomly assigned to groups of five persons for the whole session, which was conducted anonymously. A payoff table was provided to indicate possible outcomes. Depending on their decisions and the combination of players in the coalition, players received different payoffs which fell somewhere in the range of £0 to £24. They were asked to make a decision to join or not join a coalition in four 15-round treatments. Unlike in the dictator game, at the end of each round, subjects were informed about their own payoffs and the coalition formation.

The treatments were all designed to achieve a condition of dominant-strategy equilibrium. From treatment to treatment, each player had offered a strategy guiding whether s/he ought to participate in a coalition. Based on their dominant strategies, players were divided into two groups: critical and non-critical players. A critical player's dominant strategy was to cooperate and a non-critical player's dominant strategy was to stay out. As mentioned earlier, critical players who were essential to an effective coalition and noncritical players were offered different levels of free-riding incentive from an effective coalition. No critical country could be replaced, even by all noncritical players acting jointly. The setting of dominant strategy equilibrium ensures that there is a unique stable coalition structure in the corresponding coalition formation model.

In this setting, a critical player could achieve her/his highest possible payoff only when all players cooperated and a non-critical player's highest payoff could be achieved by her/him becoming the only free-rider. In addition, the lowest Pareto-efficient payoff occurred when all players took their dominant strategies.

The equitable payoff $\pi_{-i}^{e}(b_{-i})$ is the average of the highest and lowest payoffs. For critical players, the highest payoff exists if everyone participates and the worst possible payoff would be 0 if no effective coalition exists. It should be noted that the signs of reciprocal and straight kindness depend on the numerators, since the denominators are all positive numbers. A player who earns less than the equitable payoff can safely assume that some other players are hostile to her/him and, thus, becomes hostile to them in response.

For example, when all players participate in a coalition, the collective payoff reaches the highest level. The critical players have positive straight and reciprocal kindness, meaning that their welfare is greater than their monetary payoffs. Due to their lack of free-riding benefit, however, the non-critical players feel that other players are being hostile to them, and thus their welfare is lower than their monetary payoffs. In other words, non-critical players have no incentive to coordinate with others.

On the other hand, a critical player, who feels non-critical players are mean to her/him, may lead to other players undergoing costly punishment at the hands of that player. In such a case, the consequence would be to make the coalition unstable internally. When a critical player decides to take revenge through non-cooperative behavior, the coalition becomes ineffective and everyone earns nothing. In other words, all possible responses yield all players the same payoff, at which point, kindness ceases to be an issue. This situation will change only when the critical player in question believes that other players will behave cooperatively, and such a player will cooperate when s/he believes the coalition has the potential to be larger than it would be in a state of dominant-strategy equilibrium.

Together, these results provide important insights into the formation of unstable coalitions due to players' beliefs and reciprocal behavior. In a coalition of all players (also known as a 'grand coalition'), non-critical ones might feel hostile due to their non-attainment of any free-riding benefit. In a state of dominant-strategy equilibrium, on the other hand, the critical players might feel hostile toward the free-riders. Thus, coalition formation can be shaped and reshaped by the subjects' beliefs and preferences.

4. Analyses of the Experimental Evidence

Following equations (8) and (9), subjects' reciprocal kindness and straight kindness can be measured. In practice, we employ the decisions in the past to represent a player k's beliefs about the strategy of other players b_{-k} and a player k's beliefs about other players' beliefs about her strategy c_k. In other words, what players will believe and would like the others to believe are based on the past decisions made by them and their opponents. Thus, $\pi_k\left(c_k, b_{-k}\right)$ is player k's payoff in the past round, while $\pi_k\left(a_k, b_{-k},\right)$ is k's payoff given k's present and others' past decisions.

Table 1 reports descriptive statistics and data sources for the experimental results. The former includes the subjects' birth years. Altruistic attitude, determined by the dictator test, indicates the subject's average straight-kindness level across all 20 rounds of the game. Because the game did not measure reciprocal kindness, the overall average straight-kindness level was -0.21, implying that the subjects, as a group, were hostile to others across the game as a whole. Interestingly, subjects became more altruistic when the token was more valuable to receivers than to givers – showing that the value to the giver was an important factor in a subject's decision-making. Specifically, when the value of the token to the potential giver was relatively small, s/he was more likely to behave kindly by giving it up.

Table 1: Descriptive statistics and data sources

Variable	Observations	Mean	SD	Min	Max
Birth year	50	1988	4.37	1968	1992
Altruism	50	0.71	0.31	0.05	1
Membership decisions	3,000	0.68	0.47	0	1
Straight kindness	2,800	-0.35	0.45	-1.05	0.14
Reciprocal kindness	2,800	-0.35	0.45	-1.05	0.14
Reciprocal feeling	2,800	-0.05	0.18	-1.06	0.16

Table 1 also reports the subjects' decisions in the coalition game (dummy of joining or not joining a coalition). We applied this data to equations (8), (9) and (10) to illustrate straight kindness, reciprocal kindness, and reciprocal welfare. The latter comprises the subjects' shared notion of fairness, which incorporates both straight kindness and reciprocal kindness in the coalition game. As noted earlier in equations (8) and (9), the decisions made in the prior round are used to indicate players' beliefs. Hence, what the opponent players could believe indicate b_1 and b_2 by using player 2's and player 1's decisions in the past round, respectively. What the player think the opponent players would believe indicate c_1 and c_2 by using player 1's and player 2's decisions in the past round, respectively. The highest and lowest Pareto-efficient payoffs are the highest and lowest payoffs among all possible outcomes. For a critical player, the highest payoff is a grand-coalition solution, and the lowest Pareto-efficient payoff is the self-interested prediction. For a noncritical

player, the highest payoff is a solution in which the player is the only non-signatory, and the lowest Pareto-efficient payoff occurs when s/he enters a coalition that otherwise consists only of the critical players. The worst payoff occurs when no coalition is formed (i.e., everyone gets nothing). A player's straight kindness is computed as the average of her/his kindness toward all four of the other players. In the same way, by employing the players' historical decisions, we can determine a player's reciprocal kindness, i.e., subjective sense of how kind other players have been to her/him, as the average of her/his attitude to other players in the group.

When correlation coefficients were computed to assess the relationships between each individual's straight and reciprocal kindness, a significant positive correlation ($\mathbf{0.84}$) between these two constructs were identified. The mean values of straight and reciprocal kindness were $\mathbf{-0.351}$ and $\mathbf{-0.346}$, respectively, meaning that, in general, subjects were hostile to others and were treated badly by others. Average reciprocal kindness among critical players was slightly lower than among noncritical ones, i.e., $\mathbf{-0.37}$ vs. $\mathbf{-0.33}$. In general, the reciprocal kindness was negative whenever subjects were critical or noncritical. Hence, we can say that subjects behaved badly and were treated badly, in general, and that both these phenomena were more marked when they were critical players.

Due to negative reciprocal kindness, as mentioned in the previous section, the subjects would feel worse than their monetary payoffs usually provided. It is worth noting that subjects were concerned with not only about their own payoffs, but also about the payoffs received by others. Nevertheless, they felt jealous instead of proud of other's gains. The more generous they were according to the dictator game, the less likely they were to join coalitions and, thus, make contributions in the public-goods game. This interesting result may be explicable via the variable of reciprocal kindness; that is, when a subject was treated badly, s/he would be more likely to participate in the coalition. In other words, participation was not only self-enforced but could also represent compliance with a punishment method out by one or more hostile critical players.

In the membership part of the experiment, effective coalitions were formed in 387 out of 600 rounds. The structure of coalitions tended to be unstable. The size of coalitions was usually larger than the dominant strategy equilibrium, which was formed in only 112 rounds. Moreover, even within the same treatment, it varied from group to group. This implies that, even though the game was designed to favor dominant-strategy equilibrium, the formation of stable-coalition was unachievable.

As compared to the result in the first round, participation rates declined over the course of the remaining rounds, from 93% to 85% among the critical players and from 59% to 46% among the noncritical ones. This means that subjects did not look after only their self-interest but also others. However, the more they learned about other players' decisions, the less cooperative they became.

Table 2, which shows panel-data estimates of the probability-of-joining equation, covers the observations of 2,800 individual decisions in the public-goods game, the first observation in every treatment having been excluded for indicating the direct and reciprocal kindness. Amongst these observations, the subjects decided to join a coalition a total of 1,884 times.

Table 2: Panel-data estimates of the probability-of-joining equation

Variable	Pooled Least Squares	Fixed Effects	Random Effects
Constant term	0.70 (3.38)	0.52 *** (0.01)	0.63 (10.0)
(v1) Age	-0.0001 (0.002)		-0.0001 (0.01)
(v2) Altruism	-0.09 *** (0.02)		-0.09 (0.07)
(v3) Straight kindness	0.68 *** (0.03)	0.67 *** (0.03)	0.68 *** (0.03)
(v4) Reciprocal kindness	-0.50 *** (0.03)	-0.51 (0.03)	-0.51 *** (0.03)
(v5) Player role	0.40 *** (0.01)	0.39 (0.01)	0.40 *** (0.01)
R-squared	0.30	0.30	0.31
Total observations	2,800		
Observations of joining	1,884		
Hausman test	Prob>chi2 = 0.6653		
Breusch-Pagan LM test	Prob > chibar2 = 0.0000		

Note: Each cell contains coefficient and standard deviation. *** means significant at 0.5% level

Given our core interest in factors that might affect individual decisions, two time-invariant variables were included in the equation: (v1) birth year and (v2) altruism measured by the dictator game. As such, we cannot directly compare the fixed-effects and random-effects estimators, as the random-effects model provides separate estimates of the parameters on the time-invariant variables, while the fixed-effects model cannot. The variables that were subject to change between one round to another included (v3) straight kindness and (v4) reciprocal kindness, indicate the player's reciprocal preferences in the coalition game. (v5) player role is a dummy variable which describes the player's dominant strategy: 1 = critical to the coalition and her/his dominant strategy is to join, while 0 = non-critical and her/his dominant strategy is not to join.

When we performed tests for the statistical significance of the differences between the coefficient estimates obtained by the models, Hausman testing revealed that the random-effects estimates were more efficient and more consistent than the fixed-effects estimates. Additionally, the Breusch-Pagan LM test found that the random-effects estimates were more efficient than pooled least-squares ones. As such, it can be said that the individual specific effects were uncorrelated with the independent variables. However, two variables – critical player dummy and direct altruism – were positively associated with a person's probability of joining a coalition, while reciprocal altruism was negatively associated with such probability. These results are intuitive and as predicted: subjects tended to select the weakly dominant strategy, and kind subjects were more likely than others to cooperate but even more cooperative when treated badly.

As noted above, the experiment's design predetermined the number of critical players essential to form an effective coalition. Studying the behavior of critical players can enhance our understanding of their decisions, due to their role in stabilizing the coalition internally. Tables 5 and 6 break down the panel-data estimates of the probability-of-joining equation according to whether the observations were of critical or non-critical players.

Table3: only Panel-data estimates of the probability-of-joining equation (critical players)

Variable	Pooled Least Squares	Fixed Effects	Random Effects
Constant term	-1.47 (3.29)	0.94 *** (0.008)	-1.04 (10.4)
(v1) Age	0.001 (0.002)		0.001 (0.01)
(v2) Altruism	-0.06 *** (0.02)		-0.08 (0.08)
(v3) Straight kindness	0.69 *** (0.02)	0.66 *** (0.02)	0.66 *** (0.02)
(v4) Reciprocal kindness	-0.39 *** (0.02)	-0.41 *** (0.02)	-0.41 *** (0.02)
R-squared	0.41	0.40	0.40
Total observations	1,540		
Observations of joining	1,308		
Hausman test	Prob>chi2 = 0.078		
Breusch-Pagan LM test	Prob > chibar2 = 0.0000		

Note: Each cell contains coefficient and standard deviation. *** means significant at 1% level

Table 3 covers all 1,540 observed individual decisions by critical players, of which 1,308 consisted of a decision to join a coalition. The results of both Hausman and Breusch-Pagan LM testing indicated that random-effects estimates were both more efficient and more consistent than either fixed-effects or pooled least squares estimates. Again, individual specific effects were uncorrelated with the independent variables. The critical-player dummy and direct altruism were both positively and significantly associated with the probability of joining a coalition, and reciprocal altruism was negatively and significantly associated with that probability. Additionally, critical players were more likely to cooperate when they were treated badly by others. This rejects the first hypothesis: when a smaller coalition or no effective coalition had been formed in the previous round, critical players would nevertheless seek to form one in the current round.

It is worth noting that, across all treatments, the participation rate in the first round was higher than in any subsequent round. This could be explained by negative reciprocal kindness: in a coalition that is larger than dominant-strategy equilibrium would provide, critical players felt treated kindly only when all non-critical players cooperated with them, and their reactions to others became unkind when such cooperation was not forthcoming.

Table 4: Panel-data estimates of the probability-of-joining equation (non-critical players only)

Variable	Pooled Least Squares	Fixed Effects	Random Effects
Constant term	7.93 (5.50)	0.87 (0.04)	9.70 (16.8)
(v1) Age	-0.004 (0.003)		-0.004 (0.008)
(v2) Altruism	-0.07 (0.04)		-0.07 (0.12)
(v3) Straight kindness	6.03 *** (0.35)	6.32 *** (0.29)	6.29 *** (0.29)
(v4) Reciprocal kindness	-6.00 *** (0.35)	-6.32 *** (0.29)	-6.30 *** (0.29)
(v6) Marginal benefit	-2.51 *** (0.32)	-4.04 *** (0.41)	-3.69 *** (0.38)
R-squared	0.22	0.21	0.22
Total observations	1,260		
Observations of joining	576		
Hausman test	Prob>chi2 = 0.0085		
Breusch-Pagan LM test	Prob > chibar2 = 0.0000		

Note: Each cell contains coefficient and standard deviation. **, *** are significant at 5%, and 1% level, respectively

Table 4 presents the panel-data estimates for all 1,260 observations of non-critical players' individual decisions. Though all such players were offered to free-ride, the results indicate that such incentives were denied nearly half the time. Since the non-critical players had different levels

of free-riding incentive, the estimation includes the marginal benefit as (v6).

Hausman testing indicated that fixed-effects estimates were more efficient and consistent than random-effects ones, meaning that individual-specific effects were correlated with the independent variables. Specifically, straight altruism was positively and significantly associated with the probability that a person would join a coalition, whilst marginal benefit and reciprocal kindness were both negatively and significantly associated with that probability. These results are also intuitive, in the sense that higher marginal benefits of total abatement brought higher incentives, and a stronger free-riding incentive would drive them away. In contrast to the experimental results provided by Burger and Kolstad (2010), the present study found that higher marginal benefits significantly increased the coalitions' sizes. The straight kindness and reciprocal kindness had the same implications in this model as in the previous ones, meaning that the non-critical players were more likely to compromise when they felt critical players were punishing them. This supports our second hypothesis.

5. Conclusions

This study investigates the impact of individual reciprocal preference on the climate coalition formation both theoretically and experimentally. Depending on reciprocal preferences, individual welfare could be more or less than the self-interest (monetary payoff) in the reciprocal model. The theoretical predictions claim that negative reciprocal kindness could turn players' decisions away from their dominant strategies, no matter whether they are critical or not to an effective coalition. The experimental result has revealed that the coalition did not always form as the dominant strategy equilibrium. That being said, our results suggest that the decision-making process is too complex to be captured by egocentric preference. We use experimental evidence to test the hypotheses of individual decisions. When others took advantage of them, they were still more likely to cooperate. This is against the first hypothesis that negative reciprocal kindness would turn critical players away from a coalition. On the other hand, if a non-critical player felt be treated unkindly, she/he was more likely to compromise and participate in a coalition. This confirms the second hypothesis that negative reciprocal kindness leads non-critical players participating in a coalition.

In terms of policy implications, this study has shown that coalition formation could be reshaped by reciprocal preferences. Due to the strategic and complicated decision process in an interactive environment, a comprehensive investigation of factors would be required in a climate coalition in practice.

References

Bardsley, N., & Moffatt, P. G. (2007). The experimetrics of public goods: inferring motivations from contributions. Theory and Decision, 62(2), 161-193.

Barrett, S. (1994). Self-enforcing international environmental agreements. *Oxford economic papers*, 878-894.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166-193.

Bosetti, V., Carraro, C., De Cian, E., Massetti, E., & Tavoni, M. (2013). Incentives and stability of international climate coalitions: An integrated assessment. *Energy Policy*, *55*, 44-56. https://doi.org/10.1016/j.enpol.2012.12.035

Bosetti, V., Heugues, M., & Tavoni, A. (2017). Luring others into climate action: coalition formation games with threshold and spillover effects. *Oxford economic papers*, *69*(2), 410-431. doi:10.1093/oep/gpx017

Brandts, J., & Schram, A. (2001). Cooperation and noise in public goods experiments: applying the contribution function approach. *Journal of Public Economics*, *79*(2), 399-427.

Breton, M., Sbragia, L., & Zaccour, G. (2010). A dynamic model for international environmental agreements. *Environmental and Resource economics*, *45*(1), 25-48.

Burger, N. E., & Kolstad, C. D. (2010). *International Environmental Agreements: Theory Meets Experimental Evidence*.

Calzolari, G., Casari, M., & Ghidoni, R. (2018). Carbon is forever: A climate change experiment on cooperation. *Journal of Environmental Economics and Management*, *92*, 169-184. https://doi.org/10.1016/j.jeem.2018.09.002

Carraro, C. (1999). *International Environmental Agreements on Climate Change* (Vol. 13). Berlin, Germany: Springer Science & Business Media.

Carraro, C., Eyckmans, J., & Finus, M. (2006). Optimal transfers and participation decisions in international environmental agreements. *Review of International Organizations*, *1*(4), 379-396.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, *117*(3), 817-869.

d'Aspremont, C., Jacquemin, A., Gabszewicz, J. J., & Weymark, J. A. (1983). On the stability of collusive price leadership. *Canadian Journal of economics*, *1*, 17-25.

Dannenberg, A., Löschel, A., Paolacci, G., Reif, C., & Tavoni, A. (2015). On the provision of public goods with probabilistic and ambiguous thresholds. *Environmental and Resource economics*, *61*(3), 365-383.

Dickinson, D. L. (2000). Ultimatum Decision-Making: A Test of Reciprocal Kindness. *Theory and Decision*, *48*(2), 151-177. doi:10.1023/a:1005274316908

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268-298.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*(3), 817-868.

Fischbacher, U., & Gachter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review, 100*(1), 541-556. doi:10.1257/aer.100.1.541

Hadjiyiannis, C., İriş, D., & Tabakis, C. (2012). International environmental cooperation under fairness and reciprocity. *BE Journal of Economic Analysis & Policy, 12*(1).

Hahn, R., & Ritz, R. (2014). *Optimal altruism in public good provision*. Cambridge Working Papers in Economics 1403. Faculty of Economics, University of Cambridge.

İriş, D., Lee, J., & Tavoni, A. (2019). Delegation and public pressure in a threshold public goods game. *Environmental and Resource economics, 74*(3), 1331-1353. doi:10.1007/s10640-019-00371-6

Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution formation in public goods games. *American Economic Review, 99*(4), 1335-1355.

Lange, A. (2006). The impact of equity-preferences on the stability of international environmental agreements. *Environmental and Resource economics, 34*(2), 247-267.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics, 1*(3), 593-622.

Lin, Y.-H. (2017). The effects of inequality aversion on the formation of climate coalition: Theory and experimental evidence. In M. Ö. Kayalıca, S. Çağatay, & H. Mıhçı (Eds.), *Economics of International Environmental Agreements: A Critical Approach* (pp. 73-88). Abingdon, United Kingdom: Routledge.

Lin, Y.-H. (2018). How does altruism enlarge a climate coalition? *Journal of Environmental Management and Tourism, 9*(3), 553-563.

Nagashima, M., Dellink, R., Van Ierland, E., & Weikard, H.-P. (2009). Stability of international climate coalitions—a comparison of transfer schemes. *Ecological Economics, 68*(5), 1476-1487.

Nordhaus, W. (2015). Climate clubs: Overcoming free-riding in international climate policy. *American Economic Review, 105*(4), 1339-1370.

Nyborg, K. (2018). Reciprocal climate negotiators. *Journal of Environmental Economics and Management, 92*, 707-725. https://doi.org/10.1016/j.jeem.2017.08.008

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review, 83*(5), 1281-1302.

Seinen, I., & Schram, A. (2006). Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European economic review, 50*(3), 581-602.