**Hitotsubashi University**
**Department of Economics**

2-1 Naka, Kunitachi, Tokyo, Japan

Discussion Paper #2014-17

# International Cooperation and Institution Formation: A Game Theoretic Perspective

Akira Okada

December 2014

# International Cooperation and Institution Formation: A Game Theoretic Perspective [1]

Akira Okada[2]

December 2014

ABSTRACT: Game theory presents a useful analytical tool for addressing the problem of international cooperation and the formation of institutions. We first examine four problems that must be solved to achieve international cooperation: the common knowledge problem, agreement problem, compliance problem, and participation problem. An institution is a mechanism used to enforce participants to cooperate for collective benefits. We consider a multi-stage game model of institution formation and show that a group of participants voluntarily forms an institution for international cooperation in a strict subgame perfect equilibrium if and only if the group satisfies the criticality condition. Some of the implications on the international frameworks that attempt to prevent the proliferation of nuclear weapons in East Asia are finally discussed.

[1] I would like to thank Motoshi Suzuki and members of the Workshop on International Political and Economic Analysis for their valuable suggestions and comments. Financial support from the Japan Society for the Promotion of Science under Grant No. (A)26245020 is gratefully acknowledged.

[2] Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601 Japan. E-mail: *aokada@econ.hit-u.ac.jp*

# 1 Introduction

Since the publication of von Neumann and Morgenstern's (1944) monumental book *Theory of Games and Economic Behavior* 70 years ago, game theory has grown progressively as a mathematical theory to study interdependent decision making in social situations. It has not only become one of the most powerful analytical tools in such social sciences fields as economics, political science, sociology, management science, and social psychology, but has also been applied successfully to other disciplines related to natural sciences and engineering. Its widespread application to many different disciplines is driven by the commonality of its fundamental problem, namely the interdependence of multiple autonomous actors. Globally, the interdependence of states and other political actors causes various phenomena such as war, peace, conflict, cooperation, negotiations, and alliances. Theories of international relations attempt to answer the basic questions regarding these phenomena such as why war occurs, how war can be avoided, how peace is established, how states cooperate to resolve military, economic and environmental conflicts, why international institutions are needed, and how they can be created.[1]

The international world is faced with various conflicts over territory, resources, and wealth, causing political, economic, and environmental problems such as regional wars, monetary crises, free trade disputes, and global warming. To solve these problems, international cooperation and social order are needed at multiple levels. Hence, a variety of psychological, social, economic, and political mechanisms exist through which cooperation and order in society can not only emerge but also be promoted and sustained. These diverse mechanisms include friendship, reciprocity, altruism, trust, communication, learning, reputation, social norms, negotiations, agreements, institutions, and legal systems, which should work at multiple levels in a complementary fashion. In this chapter, we discuss the possibility of international cooperation from a game theoretic

[1] For an excellent survey of game theoretical works on peace and war, see O'Neill (1994).

perspective, focusing particularly on negotiations and the formation of institutions.[2]

Institutions are 'the humanly devised constraints that structure political, economic and social interaction' (North 1991). They consist of both informal constraints and formal rules. Specifically, we regard an institution as a mechanism that enforces its participants to choose a collective action. A prominent characteristic of international society is that no central government regulates states' behaviour; hence, in such an anarchic state of nature, cooperation should be voluntary. Any analysis of international cooperation must thus answer the fundamental question of how individuals seeking their own values can voluntarily create an institution that constrains their liberties. Further, cooperation should be strategically stable, that is sustained as a Nash equilibrium of a game (in game theoretic terminology).

Since the seminal publication of Luce and Raiffa (1957), a large body of research has investigated the possibility of cooperation in a dynamic situation where a game is played infinitely many times. In repeated prisoner's dilemma games, players can use the following sanction mechanism against a defector. If one player can monitor a defecting action by the other player, then he/she will never cooperate in future rounds. This sanctioning mechanism, called a trigger strategy, can deter players' deviation from cooperation if their discount rates for future payoffs are sufficiently low. The well-known result termed the folk theorem in repeated games shows that every outcome that is Pareto superior to the non-cooperative outcome can be sustained by a Nash equilibrium (and also by a subgame perfect equilibrium) in repeated prisoner's dilemma games if players are sufficiently patient (Fudenberg and Maskin 1986).

The folk theorem of a repeated game has advantages and disadvantages. Although it shows that cooperation is possible among rational and selfish players in repeated prisoner's dilemma games, its conclusion is weak in that almost every outcome is possible. This difficulty worsens in a multi-person prisoner's dilemma game in which any group of

[2] Okada (2014) reviews recent works on game theoretical analyses of cooperation and the formation of institutions.

cooperators can be formed in an equilibrium if all group members receive higher payoffs than those in the non-cooperative outcome (Taylor 1987). Moreover, the equilibrium can be sustained by a variant of trigger strategies (called a group-trigger strategy) where sanction is applied only to group members. Then, a further question arises: which cooperative group is formed and how? In this chapter, we consider the endogenous formation of cooperative groups by using multi-stage game models involving participation, implementation, and contribution.

As a case for our game model, we consider international frameworks for preventing nuclear proliferation in East Asia. North Korea joined the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) in 1985. The country, however, was reluctant to accept a safeguard agreement of the International Atomic Energy Agency (IAEA), and it was suspected of having enough plutonium to produce nuclear weapons. The end of the Cold War affected the diplomatic policies of North Korea, and it signed the safeguard cord with the IAEA in 1992. During inspections by the IAEA, inconsistencies were found between North Korea's declaration and the IAEA's findings, which suggested that the country had undeclared plutonium. The IAEA requested a special inspection in the safeguard agreements, which North Korea refused. On 12 March 1993, North Korea announced its decision to withdraw from the NPT. On 11 May 1993, the UN Security Council called upon North Korea to comply with the agreement (the first nuclear crisis). To resolve the crisis, the United States started bilateral negotiations with North Korea in 1993, and the 1994 United States–North Korea Agreed Framework was established.[3] Under this Framework, North Korea shut down its Yongbyon reactor, and the United States committed to provide LWR generating capacity through the creation of the Korean Peninsula Energy Development Organization (KEDO). South Korea and Japan participated in KEDO; however, its implementation was slower than expected. In 2002, North Korea acknowledged that it had a programme to enrich uranium for nuclear

[3] The text of the agreement can be found at http://www.iaea.org/Publications/Documents/Infcircs/Others/infcirc457.pdf (accessed 3 November 2014).

weapons and announced its withdrawal from the NPT effective as of 11 January 2003. The United States–North Korea Agreed Framework subsequently collapsed (the second nuclear crisis).

In August 2003, the Six Party Talks started multilateral negotiations involving China, the United States, North and South Korea, Japan, and Russia. The fourth round of talks in 2005 made a joint statement that 'the goal of the Six-Party Talks is the verifiable denuclearization of the Korean Peninsula in a peaceful manner'.[4] North Korea committed to abandoning all nuclear weapons and existing nuclear programmes. The United States affirmed that it had no nuclear weapons on the Korean Peninsula and no intention of attacking or invading North Korea. The joint statement, however, did not improve the situation, and the US Treasury consequently froze North Korea's accounts in Banco Delta Asia, a Chinese bank in Macau, because of its money laundering activities. North Korea refused to participate in the Six Party Talks, and on 9 October 2006, conducted its first nuclear test. The UN Security Council accepted Resolution 1718 with sanctions including an arms embargo, a ban on the export of luxury goods to North Korea, and a ban on the provision of financial services. The Six Party Talks stopped in 2008.[5]

The remainder of the chapter is organized as follows. Section 2 discusses four problems in international cooperation. Section 3 presents a game model of an international public good. Section 4 presents and analyses a multi-stage game model of the formation of institutions in the provision of international public goods. Section 5 discusses some of the implications of the model for the case of the proliferation of nuclear weapons in East Asia. All proofs are presented in the appendix.

[4]The text of the statement can be found at http://www.mofa.go.jp/region/asia-paci/n_korea/6party/joint0509.html (accessed 3 November 2014).

[5]Huntley (2007) and Wit (2007) provide strategic analyses of US policy towards North Korea after the stalemate of the Six Party Talks.

## 2 Four Problems in International Cooperation

The problem of international cooperation typically arises when providing international public goods such as UN peacekeeping, the NPT, dispute settlement by the World Trade Organization (WTO), the International Monetary Fund, the European Stability and Growth Pact, the Kyoto Protocol, and the World Health Organization. Economic theory suggests that public goods tend to be undersupplied because of the incentive to free-ride (Olson 1965). To prevent free-riding, the incentive structure in international public goods should be changed to one based on rewards and sanctions. As an example of rewards, the Official Development Assistance offered by developed countries helps developing countries join international cooperation efforts by providing financial aid and technology transfers. As examples of sanctions, the IAEA has established enforcement measures for member states that fail to comply with the NPT, while the WTO's Dispute Settlement Body (DSB) takes countermeasures against the violation of free trade agreements.

Four main problems exist in the successful provision of international public goods: the common knowledge problem, agreement problem, compliance problem, and participation problem. Each of these four problems is explained next by discussing the example of the United Nations Framework Convention on Climate Change (UNFCCC).

**(1) Common knowledge problem**

Because the global climate is affected by many uncertain variables, scientific knowledge of the causes of global warming was lacking in the early stages of international negotiations on the UNFCCC. States must perceive and understand correctly the character of international public goods. Hence, sufficient scientific knowledge is critical to their successful provision, as is the mutual understanding of their character. In game theoretic terminology, the nature of international public goods should be common knowledge among players.[6] If states do not have common knowledge about the costs and benefits

[6] An event $A$ is called common knowledge among players if everyone knows $A$, everyone knows that

of international public goods, it is hard for them to agree to provide them. International negotiation is thus a process of states improving their mutual understanding of and then obtaining common knowledge on the issues in question.

In addition to scientific knowledge, however, a state predicts the goals, intentions, and beliefs of other states based on the information available. Knowledge about these internal variables of other players is essential for making a decision; nevertheless, such an estimation about other states' conditions is genuinely subjective. States should then have a mutual understanding about each other's subjective estimations. Game theory with incomplete information (Harsanyi 1967, 1968) considers the problem of the mutual understanding of players' subjective estimations under the Bayesian hypothesis that they make probabilistic estimations on each other's uncertain variables including payoffs and beliefs in a mutually consistent manner.

### (2) Agreement problem

International negotiation is a political process where states attempt to agree to an alternative from a menu of feasible policies that are either finite or infinite. In the UNFCCC negotiations, for instance, states negotiated over the total amount of greenhouse gas emissions as well as the initial allocation of emissions for OECD countries. The Kyoto Protocol prescribed that OECD countries should aim to reduce their overall emissions by 5.2 percent compared with 1990 levels between 2008 and 2012. The reduction rates of the major countries were Russia and Ukraine 0 percent, Japan 6 percent, the United States 7 percent, and the EU 8 percent.[7] States also negotiated whether to introduce so-called Kyoto mechanisms (e.g. emissions trading, clean development, and joint implementation) to enhance the efficiency and equity of the agreement. Because an agreement is made according to a certain collective choice rule such as the unanimity and majority

---

everyone knows $A$, everyone knows that everyone knows that everyone knows $A$, $\cdots$.

[7] Okada (2003, 2007) considers an initial allocation of emissions in the Kyoto Protocol according to cooperative and non-cooperative game models.

rule, states must agree to their collective decision rule before negotiations begin.

**(3) Compliance problem**

Even if states can agree to provide international public goods jointly, the agreement should then be complied with and implemented. Do states comply with an agreement that constrains their behaviour? And if they do, what is the compliance mechanism? To guarantee the compliance of states, their incentive structures must be changed so that they do not violate the agreement. Rewards and sanctions are useful devices for this purpose. Returning to the UNFCCC case example, at the Rio de Janeiro Earth Summit in June 1992, developed countries promised to reduce greenhouse gas emissions to their 1990 levels by 2000. The UNFCCC, however, lacked any legally binding commitments and hence this voluntary approach was unsuccessful. In response, the Kyoto Protocol was agreed at COP3 in 1997 as a 'legally-binding protocol or other legal instrument'. Details of the implementation measures were discussed in later negotiations. In the Bonn Agreement[8] made at COP6-Part2 in July 2001, the parties agreed to establish a Compliance Committee with a facilitative branch and an enforcement branch. Similarly, in the Marrakech Accords[9] at COP7 in October and November 2001, non-compliant parties committed to reduce excess emissions in the subsequent commitment period by a penalty rate 30 percent. However, such a formal process of implementation involves drafting, monitoring, rewarding, sanctioning, and appealing and thus the implementation costs tend to be high. Indeed, compared with formal sanctions, Chayes and Chayes (1993) argue that an informal and managerial approach relying on social norms, persuasion, assistance, and acceptability is more effective for compliance than a formal one. Likewise, Simmons (1998) reviews different perspectives to compliance, emphasizing the use of power, law, domestic regimes, and norms.

[8] The full text of the agreement can be found at http://unfccc.int/resource/docs/cop6secpart/05.pdf.

[9] The full text of the accords can be found at http://unfccc.int/cop7/documents/accords-draft.pdf.

**(4) Participation problem**

Participation in the provision of international public goods should be voluntary. A participation rule can be either open, namely when any state can join, or closed, when the participation of a new member must be approved by the incumbents. A state may be willing to participate in the provision of international public goods if a sufficient number of states do so. Therefore, the participation decision is affected by other states' decisions on participation. In making such a complex decision, countries should reason strategically about states' decisions. Without mutual trust and sufficient incentives, few states would participate in the provision of international public goods. For example, in the UNFCCC negotiations, the United States claimed that developing countries should also agree emissions reduction targets and revealed its reluctance to participate in the Kyoto Protocol unless all developed and developing countries had individual goals. Indeed, after the Kyoto conference, the United States actually opted out of the Kyoto Protocol.

In the following, we consider cooperation in the provision of international public goods, focusing on the participation problem. For analytical simplicity, we assume that public goods are certain and that once a group of states agrees to provision, this agreement is complied with by adopting group-trigger strategies. Our game model thus aims to capture the strategic interdependence of states' participation decisions in the simplest way.

# 3 The Model

Consider the following $n$-person game $G$ of international public goods. Let $N = \{1, 2, \cdots, n\}$ be the set of countries. Every country $i \in N$ chooses its contribution $g_i$ to public goods independently. For expositional simplicity, we assume that every country $i$'s contribution is binary, either zero or a fixed value $\omega > 0$. For a contribution profile $g = (g_1, \cdots, g_n)$ of all countries, every country $i$ receives payoff

$$u_i(g_1, \cdots, g_n) = \omega - g_i + a_i \sum_{i=1}^{n} g_i, \tag{3.1}$$

where $1/n < a_i < 1$.[10] Parameter $a_i$ represents the marginal per capita return (MPCR) of country $i$ from contributing to the public goods.[11] Every country is assumed to maximize its payoff. The condition $a_i < 1$ implies that every country $i$ maximizes its payoff by contributing nothing ($g_i = 0$), regardless of the other countries' contributions, and thus that the zero contribution profile $g = (0, \cdots, 0)$ is a unique (Nash) equilibrium. The condition of $1/n < a_i$ implies that the Nash equilibrium is not Pareto optimal, that is, all countries are better off by contributing $g_i = \omega$ jointly than they are in the Nash equilibrium. In what follows, a country is called a *cooperator* if it contributes $\omega$ and a *defector* otherwise.

For a subset $S \subseteq N$, let $g^S = (g_i^S)_{i \in N}$ denote the contribution profile in which $g_i = \omega$ for every $i \in S$ and $g_i = 0$ for every $i \notin S$. Let $s$ denote the number of members in $S$. In the contribution profile $g^S$, all members of $S$ are cooperators. All countries outside $S$ are defectors, and they free-ride on cooperators' contributions. Cooperator $i$ is better off in the contribution profile $g^S$ than it is in the equilibrium $g = (0, \cdots, 0)$ if

$$a_i s \omega > \omega. \tag{3.2}$$

If (3.2) holds for every $i \in S$, a group $S$ of cooperators is called *individually rational*. In an individually rational group, all members are better off than they are in the zero contribution equilibrium.

The smallest integer $s_i^*$ satisfying (3.2) is called the *cooperation threshold* for country

[10] This game is called a linear public goods game since the supply of public goods in the payoff function (3.1) is linear in total contributions. Other types of public goods include best shot public goods and weakest link public goods. For example, Okada (2008) considers a Cobb–Douglas utility function in a public goods game with capital accumulation.

[11] Each country $i$ receives an additional payoff $a_i$ if it increases one unit of contributions.

$i$. Equivalently, $s_i^*$ is the smallest integer that is larger than $1/a_i$.[12] Since $1/n < a_i < 1$, it holds that $2 \leq s_i^* \leq n$ for every $i \in N$. Country $i$ is willing to cooperate if the number of cooperators (including itself) exceeds $s_i^*$. The cooperation threshold $s_i^*$ measures the strength of country $i$'s incentive to cooperate. Countries with smaller thresholds are more willing to cooperate than those with higher thresholds. Hence, it holds that a country with a larger MPCR has a stronger incentive to cooperate.

For two contribution profiles $g = (g_1, \cdots, g_n)$ and $g' = (g'_1, \cdots, g'_n)$, we say that $g$ is *Pareto superior* to $g'$ if $u_i(g) > u_i(g')$ for every $i \in N$. A contribution profile $g = (g_1, \cdots, g_n)$ is called *Pareto optimal* if there exists no contribution profile that is Pareto superior to it. A set $S$ of cooperators is called *Pareto optimal* if the contribution profile $g^S$ is Pareto optimal.

According to (3.1), the largest group $N$ of cooperators is always Pareto optimal since a cooperator is worse off if any country deviates from the group. It is also individually rational since $1/n < a_i$ for every $i \in N$. In general, an $n$-person game $G$ of international public goods has multiple Pareto-optimal groups of cooperators. The following proposition characterizes an individually rational group and a Pareto-optimal group of cooperators when all countries have the same MPCRs.

**Proposition 3.1.** *Suppose that all countries have the same MPCRs $a$. A group $S$ of $s$ cooperators is individually rational if and only if $\frac{1}{a} < s$. $S$ is Pareto optimal if and only if $n - \frac{1}{a} \leq s$.*

It follows from Proposition 3.1 that a group $S$ of cooperators is individually rational and Pareto optimal if and only if its number of members is greater than or equal to the threshold, which is given by the largest value of $\frac{1}{a}$ and $n - \frac{1}{a}$. The proposition can be explained as follows. By definition, a group $S$ of cooperators is individually rational if and only if $\omega < as\omega$, that is, $\frac{1}{a} < s$. All countries outside the group $S$ free-ride on the contributions of the members of $S$, and thus receive payoffs $\omega + as\omega$ according to

[12]As a regularity condition, we assume that $1/a_i$ is not an integer for any $i \in N$.

(3.1). If all $n$ countries contribute, then each of them receives payoff $an\omega$. Thus, if the group $S$ is Pareto optimal, then it must hold that $\omega + as\omega \geq an\omega$, that is, $n - \frac{1}{a} \leq s$. Conversely, if $n - \frac{1}{a} \leq s$ holds, then none of the outsiders of $S$ is better off by joining any group of cooperators. This fact means that the group $S$ is Pareto optimal. It can thus be seen from Proposition 3.1 that every individually rational group is Pareto optimal if $n - \frac{1}{a} \leq \frac{1}{a}$, i.e., $a \leq \frac{2}{n}$.

Table 3.1 shows the sizes of all individually rational groups for each value of MPCR $a$ in the interval $(\frac{1}{10}, 1)$ when $n = 10$ (Maruta and Okada 2012). As MPCR $a$ increases, the cooperation threshold, the minimum integer that is larger than $\frac{1}{a}$, decreases and thus more groups become individually rational. This fact implies that as an international public good becomes more beneficial to countries, a group of cooperators is more likely to form, even if its size is small. In this table, the integers with asterisks indicate Pareto-optimal group sizes for each value of MPCR $a$. For example, when $\frac{1}{10} < a < \frac{1}{9}$, the largest group with 10 members is a unique individually rational group and it is also Pareto optimal. When $\frac{1}{10} < a < \frac{1}{5}$, all individually rational groups are Pareto optimal. When $\frac{1}{5} < a$, some individually rational groups are not Pareto optimal. Specifically, when $\frac{1}{2} < a < 1$, all groups with at least two members are individually rational, but only groups with nine or 10 members are Pareto optimal. The contributions from only two countries are sufficient for them to become better off than they are in the zero contribution equilibrium.

| MPCR | Individually rational group sizes |
|---|---|
| $1/10 < a < 1/9$ | $10^*$ |
| $1/9 < a < 1/8$ | $9^*$ $10^*$ |
| $1/8 < a < 1/7$ | $8^*$ $9^*$ $10^*$ |
| $1/7 < a < 1/6$ | $7^*$ $8^*$ $9^*$ $10^*$ |
| $1/6 < a < 1/5$ | $6^*$ $7^*$ $8^*$ $9^*$ $10^*$ |
| $1/5 < a < 1/4$ | 5 $6^*$ $7^*$ $8^*$ $9^*$ $10^*$ |
| $1/4 < a < 1/3$ | 4 5 6 $7^*$ $8^*$ $9^*$ $10^*$ |
| $1/3 < a < 1/2$ | 3 4 5 6 7 $8^*$ $9^*$ $10^*$ |
| $1/2 < a < 1$ | 2 3 4 5 6 7 8 $9^*$ $10^*$ |

Table 3.1 Individually rational group sizes when $n = 10$

As noted earlier, all countries are free to choose their contributions to international public goods. In such an anarchic situation, the unique Nash equilibrium of zero contributions is the natural outcome of the public goods game $G$. No country has an incentive to comply with a cooperation agreement. In order to make cooperation possible, the incentive structures of countries should thus be changed. As the theory of repeated games shows, a long-term relationship is effective for this purpose through punishing defectors.

Consider a situation in which the public goods game $G$ is played over infinitely many periods. Every country $i$ has a discount factor $\delta_i$ $(0 \leq \delta_i < 1)$ for future payoffs and perfect information on the history of the game. Country $i$ maximizes the sum of discounted payoffs $\sum_{t=1}^{\infty} \delta_i^{t-1} u_{i,t}$, where $u_{i,t}$ is the payoff in period $t(= 1, 2, \cdots)$. Let $G^\infty$ be an infinitely repeated game of the $n$-person public goods game $G$. A *pure strategy* for country $i$ in the repeated game $G^\infty$ is a function that assigns a contribution level to every period depending on the history of the game before period $t$. The following proposition characterizes an equilibrium of the repeated game $G^\infty$.

**Proposition 3.2.** *Every individually rational group $S$ of cooperators can be sustained* *as a subgame perfect equilibrium of the repeated game $G^\infty$ of the n-person international public goods game G if for every $i \in S$, the discount factor $\delta_i$ of country i for the future payoff satisfies*

$$\delta_i \geq \frac{1 - a_i}{(s-1)a_i}. \tag{3.3}$$

*The zero contribution outcome is supported by a subgame perfect equilibrium of the repeated game $G^\infty$, regardless of the values of all $\delta_i$s.*

This proposition is a special case of the folk theorem in repeated games (Fudenberg and Maskin 1986). It states that every individually rational group of cooperators can be sustained because of the payoff-maximizing behaviour of countries if they are sufficiently patient. The group of cooperators adopts various strategies in the repeated game $G^\infty$. Consider the following strategy for every member in a group $S$: cooperate in the first period and keep cooperating as long as all members of $S$ do so, and contribute nothing forever otherwise. All non-members of $S$ contribute nothing in all periods. This strategy profile denoted by $\sigma^S$ is called the *group-trigger strategy* of $S$. Without loss of generality, we consider the group-trigger strategy $\sigma^S$ to be a subgame perfect equilibrium sustaining an individually rational group $S$ of cooperators. Moreover, the result of this chapter holds true for other subgame perfect equilibrium strategies such as tit-for-tat strategies. Note that punishment may be applied only to group members in $\sigma^S$. Non-members are free to defect. In what follows, we assume (3.3).

The significance of Proposition 3.2 is to prove theoretically that voluntary cooperation is possible among self-interested countries in an anarchic state of nature without a central government. However, the zero contribution outcome is also possible as a subgame perfect equilibrium of the repeated game. The proposition shows only the *possibility of cooperation*, not the necessity of cooperation. It reveals the well-known drawback of the folk theorem: a plethora of equilibria. Even if countries are expected to cooperate, it is unclear which group of cooperators is formed. Moreover, all equilibria of the repeated game $G^\infty$ except the one in which all countries cooperate involve conflict between group members and non-members since the latter free-ride on members' contributions. In summary, these results show that the participation problem is severe in international public goods.

# 4 Formation of Institutions

An institution is a rule (or mechanism) that enforces participants to cooperate for collective benefits, and may be either centralized or decentralized. A centralized institution has an authorized organization such as the IAEA in the NPT, the WTO's DSB, and the International Court of Justice, which sanctions violators of agreements. In a decentralized institution, by contrast, participants monitor each other's behaviour and sanction violators accordingly. The group-trigger strategy is an example of a decentralized institution. Any institution, either centralized or decentralized, is endowed with a certain enforcement mechanism. For example, the IAEA has various measures of safeguards[13] against violators, while the WTO settles disputes through the DSB.[14]

The institutional approach to the theory of collective action, which presumes that individuals voluntarily agree to create an enforcement institution, has been pervasively criticized in the social sciences literature. Parsons (1937:89–94) contends that Hobbes' argument on the 'Leviathan' is logically inconsistent. He questions that if individuals are rational egoists, then why should they act in the collective interest by establishing a coercive state. In his critique of 'new institutionalism', Bates (1988) also maintains

[13]According to the IAEA document, 'In the event of non-compliance and failure by the recipient State or States to take requested corrective steps within a reasonable time, [the IAEA is able] to suspend or terminate assistance and withdraw any materials and equipment made available by the Agency or a member in furtherance of the project' (ARTICLE XII:A7). Please see http://www.iaea.org/About/statute.html for more details.

[14]The aim of the DSB is to 'maintain surveillance of implementation of rulings and recommendations, and authorize suspension of concession and other obligations under the covered agreements'. (WTO Agreement, ANNEX 2, Article 2). Please see http://www.wto.org/english/docs_e/legal_e/final_e.htm.

that institutions provide a collective good and that rational individuals would seek to secure its benefits for free. The proposed institution is thus subject to 'the very incentive problems it is supposed to resolve' (Bates 1988). In other words, institutions, as second-order public goods, have the 'second-order free-rider problem' (Oliver 1980).

However, despite these negative views, Ostrom (1990) shows empirically that common-pool resources can be successfully self-governed under effective sanctioning systems. Yamagishi (1986) also presents experimental evidence that subjects voluntarily contribute to the provision of a sanctioning system in public goods games. Nevertheless, theoretical works supporting institutional approaches are scarce.

To consider the legitimacy of an institutional approach to international cooperation, we apply the multi-stage game model of institution formation provided by Kosfeld et al. (2009) to the implementation problem of group-trigger strategies. We assume that countries have preplay negotiations to implement group-trigger strategies before they actually play the repeated game $G^\infty$ introduced in Section 3. For a group $S \subset N$, let $\sigma^S$ denote the group-trigger strategy for $S$. Institution formation is modelled as the following multi-stage process.

**Institution Formation Game**

(1) (Participation stage) Every country in $N = \{1, \cdots, n\}$ announces its participation in an institution independently. The aim of the institution is to implement the group-trigger strategy. Let $S$ be the set of all participants.

(2) (Implementation stage) All participants either accept or reject the implementation of the group-trigger strategy $\sigma^S$ independently. The strategy $\sigma^S$ is implemented if and only if all participants accept it. A group $S$ is formed if all members accept the group-trigger strategy $\sigma^S$.

(3) (Contribution stage) All countries play the repeated game $G^\infty$. If a group $S$ is formed with the agreement of $\sigma^S$ in the previous stage, then all member countries

in $S$ choose their contributions according to the group-trigger strategy $\sigma^S$. All non-members make zero contributions in every period. Otherwise, all $n$ countries make zero contributions in all periods.

We call this game the *institution formation game* and denote it by $\Gamma$. In the game $\Gamma$, every country chooses its action based on its knowledge of other players' actions in all previous stages. If a group $S$ of participants implements the group-trigger strategy $\sigma^S$, then every member $i \in S$ receives the per-period payoff

$$u_i = a_i s\omega \tag{4.1}$$

and every non-member $j \notin S$ receives the per-period payoff

$$u_j = \omega + a_j s\omega. \tag{4.2}$$

Note that all members $i$ of $S$ contribute $g_i = \omega$ in all periods in the group-trigger strategy $\sigma^S$ and that all non-members $j$ contribute $g_j = 0$ in all periods. If the group $S$ does not implement $\sigma^S$, then no country $i \in N$ contributes to the public goods and thus every country receives the per-period payoff $\omega$.

The institution formation game is formulated to capture some of the basic elements of the actual formation of institutions under international cooperation. Several remarks are in order. First, participation in the institution should be voluntary. No central government exists to compel any country to participate in international cooperation. The game starts with the participation stage in which all countries voluntarily decide to participate in an institution. An institution is implemented according to the unanimity rule within the set of participants. Every participant has the veto in the implementation of the institution. Second, the process is dynamic. Countries make their decisions sequentially, updating their expectations about the prosperity of an institution based on others' behaviour. They can decide whether to accept the institution, knowing the number of participants. If few countries participate in an institution, then countries with

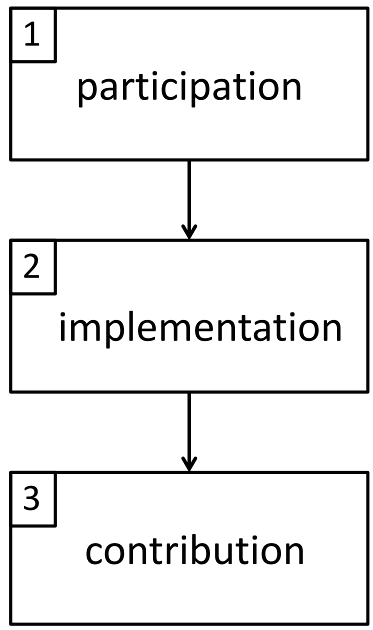


**Figure 4.1** Three stages in the institution formation game

high cooperation thresholds are likely to reject the institution and thus it fails. Third, for analytical simplicity, we ignore any costs of institution formation (e.g. those related to communication, negotiations, monitoring, punishment, staffing, and maintenance). Our analysis can be applied to the case of costly institutions without any critical change. Finally, we assume that the preplay stage of negotiations for institution formation is the only opportunity for countries to coordinate their strategies in the repeated game $G^\infty$. If negotiations fail, then they play the one-shot Nash equilibrium of zero contributions repeatedly. Figure 4.1 illustrates the sequence of three stages in the institution formation game.

We consider a subgame perfect equilibrium of the institution formation game $\Gamma$. For simplicity, we restrict our analysis to an equilibrium in pure strategies.[15] The equilibrium can be computed by using backward induction. Since the play of the contribution stage is automatically determined by the rule of $\Gamma$, depending on whether an institution is formed in the implementation stage, we proceed to analyse the second stage of implementation. The following proposition presents the result of this implementation stage.

**Proposition 4.1.** *Let $S \subset N$ be a set of participants.*
*(1) If $S$ is an individually rational group, then the implementation stage has a Nash equilibrium where all participants accept the implementation of the institution. There are also other Nash equilibria where the institution is not implemented.*
*(2) If $S$ is not an individually rational group, then the institution is not implemented in any Nash equilibrium of the implementation stage.*

The proposition can be explained intuitively as follows. If a set $S$ of participants is individually rational, then all participants are better off by implementing the institution than they are in the zero contribution equilibrium. Under the unanimous rule, the institution is not implemented if any one participant rejects it. Thus, it is a Nash equilibrium of the implementation stage that all participants accept the institution. There are, however, other Nash equilibria where at least two participants reject the institution. In these Nash equilibria, the decision of no single participant changes the outcome under the unanimous rule.

Given the equilibrium outcomes of the implementation stage, we finally analyse the participation stage. Two types of subgame perfect equilibria exist in the institution formation game $\Gamma$. A subgame perfect equilibrium of $\Gamma$ is called an *institutional equilibrium* if an institution is implemented on the equilibrium play. A subgame perfect equilibrium of $\Gamma$ is called a *status-quo equilibrium* if any institution is not implemented on the equilibrium play.

[15]Okada (1993) considers a mixed strategy equilibrium in the institution formation game.

**Proposition 4.2.** *In the institution formation game $\Gamma$, there exists an institutional equilibrium where an institution $S$ is implemented if and only if $S$ is individually rational. For every integer $s$ ($1 \leq s \leq n$), there exists a status-quo equilibrium where $s$ countries participate in an institution, which fails.*

Contrary to the negative view of an institution explained earlier, Proposition 4.2 shows that it can be voluntarily formed in a subgame prefect equilibrium of the institution formation game $\Gamma$. The institution formation game, however, has as many equilibrium outcomes as the repeated game $G^{\infty}$ in the folk theorem (Proposition 3.2). Every individually rational group $S$ can be formed in an equilibrium. Furthermore, there always exists the status-quo equilibrium with no institution. The multiplicity of equilibria in the institution formation game $\Gamma$ is caused by the unanimous voting rule of the implementation stage, which has an equilibrium where an institution is rejected (Proposition 4.1). This equilibrium, called the *rejection equilibrium*, plays a role in punishing non-participants. If the equilibrium number of countries do not participate in an institution (off the equilibrium play), then participants reject the formation of their groups. An institution of an individually rational group $S$ can thus be supported by the following strategies: all members of $S$ participate in an institution, and they accept the implementation of the institution on the equilibrium play. If a set of participants is not equal to $S$ (off the equilibrium play), then the institution is rejected in the implementation stage. The status-quo equilibrium prevails if all participants reject an institution, regardless of how many countries participate in it.

In the institution formation game, the rejection equilibrium of the unanimous voting game (the implementation stage), which causes the multiplicity of subgame equilibria, has the following unreasonable properties. First, countries rejecting an institution are indifferent to their decisions. Second, if a set of participants is individually rational, then the action of rejection is weakly dominated by that of acceptance for every participant. In other words, every participating country is (weakly) better off by accepting an institution than it is rejecting it, whichever actions the other participants choose. Specifically, it is strictly better off by accepting if all other participants also accept the institution. The rejection equilibrium can be eliminated by using an equilibrium refinement called a *strict* Nash equilibrium where every player has a unique best response to all other players' actions. A subgame perfect equilibrium of the institution formation game $\Gamma$ is called *strict* if it induces a strict Nash equilibrium on every stage game, both on and off equilibrium plays.

We next characterize a strict subgame perfect equilibrium of the institution formation game $\Gamma$. The following notion is important in the analysis.

**Definition 4.1.** For an individually rational group $S$ of $N$, country $i \in S$ is *critical* to $S$ if the group $S - \{i\}$ is not individually rational. An individually rational group $S$ is *critical* if all countries of $S$ are critical to $S$.

The notion of a critical member in the group $S$ has the following meaning. It follows from Proposition 4.1 that if the group $S$ is individually rational, then $S$ is formed in a Nash equilibrium of the implementation stage. If a critical member $i$ opts out of $S$, then the remaining group $S - \{i\}$ does not form in an equilibrium since it is no longer individually rational. Thus, country $i$ is considered to be critical to the formation of $S$. A group is called critical if every member is critical to its formation. A critical group $S$ is not sustained if any single member opts out.

**Proposition 4.3.** *In the institution formation game $\Gamma$, there exists a strict subgame perfect equilibrium where an institution with a set of participants $S$ is formed if and only if $S$ is a critical group. The status-quo equilibrium is not strict.*

This proposition shows that the equilibrium refinement of strictness predicts that only a critical group is formed in the institution formation game. The intuition behind this result can be explained as follows. The property of strictness can eliminate the rejection equilibrium in the implementation stage. In a strict subgame perfect equilibrium, every individually rational group is accepted to form in the implementation stage. If a group $S$ is not critical, and thus if there exists at least one member country that is not critical to it, then such a country has an incentive to opt out of the group because the smaller group without it still forms. Thus, the group $S$ is not supported in the equilibrium. On the contrary, every critical group can be supported by a strict subgame perfect equilibrium because any participant, being critical to the group, is worse off by deviating from it. In the status-quo equilibrium, every non-participant is indifferent to its participation decision. Thus, the status-quo equilibrium does not induce a strict Nash equilibrium in the participation stage.

A critical group can be characterized in terms of the cooperation thresholds for countries. When all countries have the same cooperation thresholds $s^*$, the number of members in a critical group is equal to $s^*$ and thus it is the smallest individually rational group. Table 2.1 shows the size of a critical group in the case of $n = 10$, depending on every possible value of MPCR $a$. The following proposition characterizes a critical group in a general case that countries have different cooperation thresholds.

**Proposition 4.4.** *A group $S$ with $s$ countries is critical if and only if (1) the group size $s$ is equal to or greater than the cooperation threshold $s_i^*$ of every country $i \in S$ and (2) there exist at least two members $i \in S$ such that $s_i^* = s$.*

Proposition 4.4 can be proved as follows. Condition (1) is a necessary and sufficient condition for the group $S$ to be individually rational. The necessity of condition (2) can be explained as follows. Suppose that condition (1) holds. If there exists only one member $i \in S$ such that $s_i^* = s$, then member $i$ is not critical to $S$ since the group $S - \{i\}$ is individually rational. If there exists no member $i \in S$ such that $s_i^* = s$, then the remaining group $S - \{i\}$ after every member $i \in S$ opts out is still individually

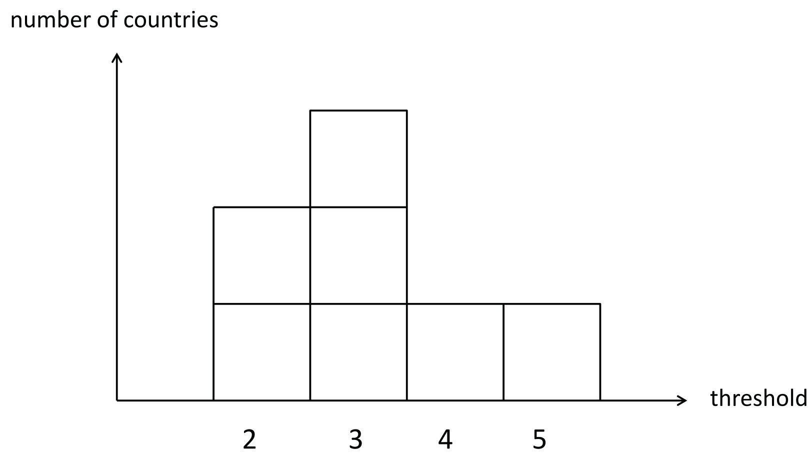


**Figure 4.2** Distribution of cooperation thresholds in a seven-country game

rational. This fact means that no member of $S$ is critical to it. Thus, if $S$ is a critical group, then condition (2) must be true. Conversely, if condition (2) holds, then every member $i$ of $S$ is critical to it since $S - \{i\}$ is not individually rational.

Figure 4.2 helps us explain how the distribution of countries' cooperation thresholds determines a critical group that is sustained in a strict subgame perfect equilibrium of the institution formation game. In this figure, there are seven countries, each of which corresponds to a small box. The cooperation threshold of two countries is two, three countries have a cooperation threshold of three, and one country each has a threshold of four and five. According to Proposition 4.4, three types of critical groups exist: one of two countries (both with a cooperation threshold of two) and two other ones of three countries (three countries with a threshold of three, and one country with a threshold of two and two countries with a threshold of three).

By using the following example, we can analyse the institution formation game in the case of three countries.

**Example.** Three-country institution formation game

Consider a three-country public goods game. Three countries, 1, 2, and 3, have the same endowment $\omega = 10$, but different MPCRs $a_1 = 0.7, a_2 = 0.6, a_3 = 0.4$, respectively. The cooperation thresholds for countries are given by $s_1^* = s_2^* = 2$, and $s_3^* = 3$. Countries 1 and 2 have higher incentives to cooperate than country 3. Two groups, $\{1, 2\}$ and $\{1, 2, 3\}$, are individually rational, and only $\{1, 2\}$ is a critical group.

Table 4.1 illustrates the payoff matrix of the public goods game. For each $i = 1, 2, 3$, actions $C_i$ and $D_i$ mean the full contribution and the zero contribution of country $i$, respectively. $D_i$ is the dominant action of every country $i$ and thus the game has a unique Nash equilibrium $(D_1, D_2, D_3)$. In the equilibrium, no country contributes to public goods.

$C_3$:

| | $C_2$ | $D_2$ |
|---|---|---|
| $C_1$ | 21, 18, 12 | 14, 22, 8 |
| $D_1$ | 24, 12, 8 | 17, 16, 4 |

$D_3$:

| | $C_2$ | $D_2$ |
|---|---|---|
| $C_1$ | 14, 12, 18 | 7, 16, 14 |
| $D_1$ | 17, 6, 14 | 10, 10, 10 |

**Table 4.1** Three-country public goods game

The participation stage of the institution formation game has the payoff matrix in Table 4.2 under the assumption that every individually rational group is accepted to form in a strict Nash equilibrium of the implementation stage. In Table 4.2, action $P_i$ means the participation of country $i$ and action $NP_i$ means non-participation. Individually rational groups $\{1, 2\}$ and $\{1, 2, 3\}$ form. The participation stage has a unique strict Nash equilibrium $(P_1, P_2, NP_3)$ in which only countries 1 and 2 participate in an institution. The largest group $\{1, 2, 3\}$ is not stable since country 3 with the lowest MPCR is reluctant to cooperate and has an incentive to opt out of the group. Note that no country is worse off than it is in the zero contribution equilibrium in Table 4.2, owing to the choice of non-participation.

$P_3$

| | $P_2$ | $NP_2$ |
|---|---|---|
| $P_1$ | 21, 18, 12 | 10, 10, 10 |
| $NP_1$ | 10, 10, 10 | 10, 10, 10 |

$NP_3$

| | $P_2$ | $NP_2$ |
|---|---|---|
| $P_1$ | 14, 12, 18 | 10, 10, 10 |
| $NP_1$ | 10, 10, 10 | 10, 10, 10 |

**Table 4.2** Three-country institution formation game

# 5 Discussion

In this section, we discuss some of the implications of our formal model of institution formation for the case of the international frameworks that attempt to regulate the proliferation of nuclear weapons in East Asia. We first examine how the model can capture historical negotiations in order to prevent nuclear proliferation in North Korea, as discussed in the Introduction.

In the case of the 1994 United States–North Korea Agreed Framework, bilateral negotiations between the United States and North Korea took place. In addition, South Korea and Japan participated in KEDO. Hence, it is appropriate to presume that these four counties participated in the negotiations on the formation of an institution (the first stage in Figure 4.1) and then agreed to the Framework (the second stage in Figure 4.1). According to the Agreement, the United States provided North Korea with an LWR project with a total generating capacity of approximately 2,000 MW(E) as well as heavy oil for heating and electricity production. In return, North Korea froze its graphite-moderated reactors and related facilities. As described in the Introduction, however, the Agreement was not complied with (the third stage in Figure 4.1). In summary, the 1994 Agreed Framework was successful in the first two stages of the institution formation process in Figure 4.1, but failed in the last stage.

In the second case of the Six Party Talks, six countries—China, the United States, North and South Korea, Japan, and Russia—participated in negotiations (the first stage in Figure 4.1). These countries had six rounds of talks in Beijing in 2003–2008 (the second stage in Figure 4.1). China played a leading role as the host country. The fourth round of talks achieved an important milestone with the joint statement that all six parties unanimously reaffirmed that the goal of the talks was the verifiable denuclearization of the Korean Peninsula in a peaceful manner. However, after North Korea conducted its first nuclear test in 2006, the talks reached stalemate. The countries thus failed to make an agreement in the second stage, and the game did not go to the third stage. In summary, although the six countries participated in the first stage in Figure 4.1, the Six Party Talks failed in the second stage.

From our formal analysis of institution formation, we can obtain the condition for the six countries to agree to establish an institutional framework preventing the proliferation of nuclear weapons in East Asia. This condition consists of the following two parts.

(1) (Acceptability) The institutional framework should be beneficial to all countries.

(2) (Criticality) Every country should be critical to the formation of the framework in the sense that the framework fails without its participation.

Acceptability is a necessary and sufficient condition for the framework to be agreed in the implementation stage. Criticality is a necessary and sufficient condition for an equilibrium of the participation stage to exist, in which all countries participate in the framework. If the criticality condition fails, then at least one country has an incentive not to participate. In the case of the 1994 United States–North Korea Agreed Framework, the criticality condition was not satisfied for China and Russia. These countries were thus not motivated to participate in the Framework. Similarly, in the case of the Six Party Talks, the acceptability condition was not satisfied.

To scrutinize the conditions for the case in question more in detail, we extend the payoff function (3.1) of a country in the international public goods game to describe more realistic conditions. Because the compliance mechanism was insufficient, we can assume that there exists some probability that North Korea may violate the agreement. Let $q$ denote this violation probability. If violation occurs, all countries $i$ other than North Korea receive a large loss denoted by $L_i$ and incur institutional costs $C_i$. In this extended set-up, the expected payoff function of every country $i$ except North Korea in the framework is given by

$$Eu_i = a_i(s - 1 + 1 - q)\omega - C_i - qL_i,$$

where $qL_i$ is the expected loss because of the possibility of non-compliance and $s$ is the number of participants. Then, the profitability condition (3.2) of an institution with $s$ participants is modified as

$$a_i(s - q)\omega - C_i - qL_i \geq \omega.$$

This set-up implies that the framework that aims to regulate the proliferation of nuclear weapons is beneficial to country $i$ if

$$s \geq \frac{1}{a_i}(1 + \frac{C_i + qL_i}{\omega}) + q.$$

In this extended model, the cooperation threshold $s_i^*$ for country $i$ is given by

$$s_i^* = \text{the smallest integer greater than } \frac{1}{a_i}(1 + \frac{C_i + qL_i}{\omega}) + q.$$

According to Proposition 4.4, the cooperation thresholds for all countries must be six for the Six Party Talks to succeed. This fact means that the sum of institutional costs and expected losses, $C_i + qL_i$, has a certain upper bound for all countries except North Korea. If the institutional costs and expected losses are very high, then the framework is never beneficial for these countries. Interestingly, it must also hold that the sum of institutional costs and expected losses has a lower bound, otherwise at least one country would have a threshold below six and would be motivated to opt out the framework. For North Korea to participate in the framework, it must stand to benefit from it.

# 6 Conclusion

In this chapter, we considered the problem of international cooperation and the formation of institutions from a game theoretic perspective. We discussed four problems that must be solved successfully to achieve international cooperation. Among them, the participation problem is critical in the current global environment (i.e. a world without a central government). In other words, countries have an incentive to free-ride on the provision of international public goods. To consider the participation problem, we presented a three-stage game model of institution formation. We proved that a group of countries voluntarily participates in an institution and implements it if and only if every member is critical to its formation and discussed some of the implications of this model for international frameworks that attempt to regulate the proliferation of nuclear weapons in East Asia. In conclusion, game theory provides a useful analytical methodology for solving the problem of international cooperation and the formation of institutions.

# 7 Appendix

### Proof of Proposition 3.1

The first part of the proposition can be easily seen by the definition of individual rationality. Suppose that a group $S$ of cooperators is Pareto optimal. Then, the largest group $N$ of cooperators is not Pareto superior to $S$. If $S = N$, then the condition $n - \frac{1}{a} \leq s$ trivially holds. If $S \neq N$, then a defector outside $S$ is no better off when $N$ is formed than when $S$ is formed. This finding implies that $\omega + as\omega \geq an\omega$, that is, $n - \frac{1}{a} \leq s$. Conversely, suppose that $S$ is not Pareto optimal. Then, there exists some group $T \neq S$ such that $T$ is Pareto superior to $S$. Then, there exists some $i \in T - S$. Otherwise, $T$ is never Pareto superior to $S$ according to (3.2). By supposition, it holds that $\omega + as\omega < at\omega$. Since $at\omega \leq an\omega$, it holds that $\omega + as\omega < an\omega$, that is, $n - \frac{1}{a} > s$. This is a contradiction. Q.E.D.

Proposition 3.1 can be generalized to the case that countries have different MPCRs. A group $S$ of cooperators is Pareto optimal if and only if for every $T$ with $T - S \neq \emptyset$ there exists some $i \in T - S$ such that $s \geq t - \frac{1}{a_i}$.

**Proof of Proposition 3.2**

For an individually rational group $S$, consider the group-trigger strategy such that all members of $S$ cooperate in the first period and keep cooperating as long as all members of $S$ do so (otherwise they contribute nothing forever). All the non-members of $S$ contribute nothing in all periods. The discounted payoff of every member of $S$ for this strategy is thus $\frac{a_i s\omega}{1-\delta}$. If a member defects, then its discounted payoff is at most $\omega + a_i(s-1)\omega + \frac{\delta}{1-\delta}\omega$. Thus, the group-trigger strategy for $S$ is a subgame perfect equilibrium of the repeated game $G^\infty$ if $\frac{a_i s\omega}{1-\delta} \geq \omega + a_i(s-1)\omega + \frac{\delta}{1-\delta}\omega$. This condition is equivalent to (3.3). Q.E.D.

**Proof of Proposition 4.1**

(1) Suppose that a set of participants $S$ is individually rational. If all participants $i \in S$ accept the implementation of $S$, then they receive the per-period payoff $a_i s\omega$. Otherwise, the institution is rejected and thus they receive the per-period payoff $\omega$. Since $a_i s\omega > \omega$ for all $i \in S$, the action profile that all participants accept the implementation of $S$ is a Nash equilibrium of the implementation stage. According to the rule of unanimous

voting, the action profiles that at least two participants reject the implementation of $S$ are also Nash equilibria.

(2) Suppose that a set of participants $S$ is not individually rational. Then, there exists at least one participant that is strictly better off by rejecting the institution $S$ than when $S$ is implemented. Note that the case of $a_i s\omega = \omega$ is omitted for all $i \in S$ since $\frac{1}{a_i}$ is not an integer. Thus, the action profile that all participants accept the implementation of $S$ is not a Nash equilibrium of the implementation stage. Q.E.D.

**Proof of Proposition 4.2**

If an institution $S$ is implemented in a subgame perfect equilibrium, then $S$ must be individually rational from Proposition 3.1. Conversely, let $S$ be an individually rational group. Consider the following strategy profile. All members of $S$ participate in an institution and no other countries participate. In the implementation stage, $S$ is accepted and all other sets of participants are rejected. It can be shown without much difficulty that this strategy profile induces a Nash equilibrium in the participation stage. For any number of participants $s$ ($1 \leq s \leq n$), consider the strategy profile that all participants reject an institution, regardless of the number of participants. This strategy trivially composes the status-quo equilibrium. Q.E.D.

**Proof of Proposition 4.3**

According to Proposition 4.1, every individually rational group is accepted to form in a strict subgame perfect equilibrium. Given this equilibrium outcome of the implementation stage, it is sufficient to prove that an action profile of the participation stage is a strict Nash equilibrium if and only if a set of participants $S$ is a critical group. First, suppose that $S$ is a critical group. If any member $i$ of $S$ does not participate in $S$, then $S - \{i\}$ is not individually rational and thus it is not implemented. Country $i$ is worse off by deviating from $S$ and thus the action profile with $S$ is a strict Nash equilibrium. Second, suppose that $S$ is individually rational but not critical. Then, there exists some member $i$ of $S$ that is not critical to $S$. If country $i$ deviates from $S$, then it is better off because $S - \{i\}$ is individually rational and is implemented. This finding means that the action profile with $S$ is not a Nash equilibrium. Finally, suppose that $S$ is not individually rational. Then, $S$ is not implemented according to Proposition 4.1. Choose any $j \notin S$. Country $j$ is better off by joining $S$ if $S \cup \{j\}$ is implemented. Otherwise, the per-period payoff of country $j$ does not change. This means that the action profile with $S$ is not a strict Nash equilibrium. Q.E.D.

# References

Bates RH (1988) Contra contractarianism: some reflections on the new institutionalism. Politics and Society 16: 387-401

Chayes A, Chayes AH (1993) On compliance. International Organization 47: 175-205

Fudenberg D, Maskin E (1986) The folk theorem in repeated games with discounting or with incomplete information. Econometrica 54: 533–554

Harsanyi JC (1967, 1968) Games with incomplete information played by 'Bayesian' players, parts I-III. Management Science 14: 159-182, 320–334 and 486-502

Huntley, WL (2007) U.S. policy toward north Korea in strategic context. Asian Survey 47: 455-480.

Kosfeld M, Okada A, Riedl A (2009) Institution formation in public goods games. Amer. Econ. Rev. 99: 1335-55

Luce, RD, Raiffa H (1957) Games and decisions: introduction and critical survey. Wiley, New York

Maruta T, Okada A (2012) Dynamic group formation in the repeated prisoner's dilemma. Games and Economic Behavior 74: 269-284

North, DC (1991) Institutions. Journal of Economic Perspectives 5: 97-112

Okada A (1993) The possibility of cooperation in an n-person prisoners' dilemma with institutional arrangements. Public Choice 77: 629-656

Okada A (2003) A market game analysis of international CO2 emissions trading: evaluating initial allocation rules. In: Sawa T (ed), International frameworks and technological strategies to prevent climate change. Springer, Tokyo, pp 3-21

Okada A (2007) International negotiations on climate change: a noncooperative game analysis of the Kyoto protocol. In: Avenhaus R and Zartman IW (eds.), Diplomacy games. Springer, Berlin, pp 231-250

Okada A (2008) The second-order dilemma of public goods and capital accumulation. Public Choice 135: 165-182

Okada A (2014) Cooperation and institution in games. Discussion paper No. 2014-11, Department of Economics, Hitotsubashi University

Oliver P (1980) Rewards and punishments as selective incentives for collective action: theoretical investigations. American Journal of Sociology 85: 1356-1375

Olson M (1965) The logic of collective action: public goods and the theory of groups. Harvard University Press, Cambridge, MA

O'Neill B (1994) Game theory models of peace and war. In: Aumann, RJ and Hart, S (eds.), Handbook of game theory with economic applications, vol 2, North-Holland, Amsterdam, pp 995-1053

Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge.

Parsons T (1937) The structure of social action. MacGraw-Hill, New York

Simmons BA (1998) Compliance with international agreements. Annual Review of Political Science 1: 75-93

Taylor M (1987) The possibility of cooperation. Cambridge University Press, Cambridge, MA von Neumann J and Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton

Wit, JS (2007) Enhancing U.S. engagement with north Korea. The Washington Quarterly 30: 53-69.

Yamagishi T (1986) The provision of a sanctioning system as a public good. Journal of Personality and Social Psychology 51: 110-116