An Intention Signaling Strategy for Indirect Reciprocity: Theoretical and Empirical Studies

Tanaka, Hiroki

(Degree)
博士（学術）

(Date of Degree)
2017-03-25

(Date of Publication)
2018-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第6800号

(URL)
https://hdl.handle.net/20.500.14094/D1006800

December 8, 2016

The doctoral dissertation

An Intention Signaling Strategy for Indirect Reciprocity: Theoretical and Empirical Studies

（間接互恵状況における意図シグナル戦略：理論・実証研究）

Graduate School of Humanities

The Division of Human Social Dynamics

122L012L

Hiroki Tanaka

Abstract

Unlike many other species, human beings cooperate even when they do not expect direct reciprocation. Indirect reciprocity (Alexander, 1987; Nowak & Sigmund, 2005) is an evolutionary explanation for this type of cooperation: A helps B, then C (someone other than B) helps A when she/he is in need. However, in order to maintain a cooperative equilibrium, the system of indirect reciprocity has to solve a difficult problem: Two types of defection (i.e. defection by free-riders who refuse to help everyone and defection by cooperative players who selectively defect on free-riders) need to be distinguished. Although this problem can be solved if cooperative players take into account second-order reputation information (i.e. a current partner's previous partner's reputation), empirical evidence concerning whether people readily utilize such information is mixed. Therefore, in the present study, I proposed intention signaling strategy (*intSIG*). *IntSIG* allows apparent defectors (who selectively defect on other non-cooperative players) to protect their reputations by abandoning some resource. Hence, *intSIG* depends on defector's voluntary communication of intention, as well as on the intention-reading ability of interaction partners. Evolutionary game analyses and a series of computer simulations support the theoretically validity of *intSIG* as a strategy for the evolution of indirect reciprocity. Furthermore, two experiments showed that people behaved in an *intSIG*-like manner. In sum, the present research provides both theoretical and empirical support for this strategy. These results underscore the importance of intention signaling in human cooperation.

Key words: indirect reciprocity, reputation, costly signaling, intention

Contents

Chapter 1

1. Introduction

There are many cooperative species in the world. Eusocial species, such as wasp or honeybees, cooperate with their blood relatives and form a well-hierarchized society (Batra, 1968; Crespi & Yanega, 1995; Michener, 1969). Some birds even take care of their relative chicks that are not their own offspring (Brown, 1978; Hatchwell et al., 2004; Hatchwell & Sharp, 2006). Vampire bats give blood sucked from livestock to their starving allies (Wilkinson, 1984, 1988). Primates form a stable bond with a specific partner through mutual grooming (Seyfarth & Cheney, 1984). Despite these rich instances, it can be said that human beings are quite distinct from other species in their ability to cooperate with others. The reason is that a system of our cooperation is amenable to not only explanations that are applicable in other species' behavior, but also a human-specific principle.

Biologically, cooperation is defined as incurring cost to confer benefit on others (see Nowak, 2012). In evolutionary biology, a behavior that decreases an actor's *fitness* (an average number of offspring), while increasing a target's fitness is regarded as cooperation. As natural selection is a process to weed out individuals with low fitness, such a "wasteful" behavior is supposed to be selected out. This logic would seem to lead to selfish (or free-rider) organisms who save the cost of cooperation to easily dominate a population at the expense of altruists because of their frugality. To see the actual world, however, cooperation is ubiquitous, as mentioned previously. It means that there must

be some biological principles solving this paradox: Why does cooperative behavior exist?

The puzzle of the evolution of cooperation has long attracted many great minds' attention. Kin selection theory, which was proposed by Hamilton (1964), was a major breakthrough for this puzzle. According to this theory, altruistic behavior toward a genetically related individual not only reduces the actor's fitness, but also indirectly enhances his/her fitness by increasing the fitness of the target, who probabilistically shares the same genes with the actor. In other words, helping a kin member is equivalent to probabilistically helping one's own genes. Therefore, one's net fitness (technically *inclusive fitness*) is determined by the direct cost of the altruistic behavior and the indirect benefit accruing from it. A highly cooperative community of honeybees and helper birds' behaviors can be accounted for by this principle. On the other hand, cooperation beyond relatives, such as the vampire bats' blood sharing and primates' mutual grooming, cannot be explained by kin selection theory. Instead, cooperation within a stable partnership, regardless of partners' relatedness, is evolvable by direct reciprocity, whereby the cost of helping a partner is compensated by the benefit of being helped by the partner (Trivers, 1971). Axelrod (1984) formalized this notion as the tit-for-tat strategy (TFT), in which one cooperates with a partner if she/he cooperated previously and refuses to cooperate if she/he refused to cooperate (Axelrod & Hamilton, 1981; Axelrod, 1984).

Unlike other species, the cooperation of human beings appears beyond kinship and stable dyadic partnership (Bowles & Gintis, 2011; Fehr & Fischbacher, 2003;

Nowak & Highfield, 2011). We cooperate even with someone whom we have never seen before or whom we do not expect to see again, which cannot be explained by kin selection theory and direct reciprocity. Taking someone's lost wallet to a police station, donating to poor people who live in a remote country, and engaging in the costly reduction of greenhouse gas emissions for future generations are pervasive in our society, while no other species have ever been observed to exhibit this level of cooperation. Such human-specific cooperation is also haunted by the adaptive problem of free-riders; hence, a central purpose of present study is to provide an evolutionarily plausible explanation for this behavior.

1.1. Problems with previous models of reputation-based cooperation

(a) All defectors are not necessarily "bad"

Even if the cost of cooperation is not recouped by the partner's reciprocal cooperation, altruists may receive the cooperation of *someone else* because of their good reputation, while free-riders may not be chosen as a target of cooperation because of their bad reputation. *Indirect reciprocity* (Alexander, 1987; Nowak & Sigmund, 2005) is a system of cooperation based on this kind of reputation dynamics—A helps B, then not B, but C helps A when he/she is in need. This system is implied in a proverb "one good turn deserves another," Accordingly, this reputation-based cooperation possibly enables humans to achieve and maintain large-scale cooperation, which is beyond the scope of direct reciprocity. Note that this system does not require people to always consciously calculate the benefit of acquiring a good reputation. In fact, we often behave

altruistically out of unconscious factors, such as emotions. Indirect reciprocity does not explain the underlying psychological mechanism of human-specific cooperation but it does explain why such form of cooperation evolved (see Tinbergen, 1963).

Nowak and Sigmund (1998a, b) first mathematically formalized this concept. They showed that cooperative equilibrium is maintained without dyadic reciprocation if individuals use the *image-scoring strategy (IS),* in which a person selectively bestows a "good" reputation on cooperative individuals and cooperates only with individuals with a good reputation (as noted previously, the word "strategy" does not imply any conscious reasoning). In fact, participants without any knowledge of game theory behaved in an *IS*-like manner in experimental games (Bolton, Katok, & Ockenfels, 2004, 2005; Dufwenberg, Gneezy, Güth, & Van Damme, 2001; Engelmann & Fischbacher, 2009; Ernest-Jones, Nettle, & Bateson, 2011; Jacquet, Hauert, Traulsen, & Milinski, 2012; Manfred Milinski, Semmann, & Krambeck, 2002a, b; Rockenbach & Milinski, 2006; Seinen & Schram, 2006; Sommerfeld, Krambeck, Semmann, & Milinski, 2007; Wedekind & Braithwaite, 2002; Wedekind & Milinski, 2000; Yoeli, Hoffman, Rand, & Nowak, 2013). Furthermore, this tendency was observed even among preschool children (Kato-Shimizu, Onishi, Kanazawa, & Hinobayashi, 2013).

However, the *IS* has a theoretical flaw that any defection (in terms of game theory, "defection" means to "not helping others") derived from an *IS*-like manner cannot be justified by *IS* players themselves (Leimar & Hammerstein, 2001). Suppose that you encounter a person who has a "bad" reputation (i.e., a free-rider). If you are an *IS* player, you will defect on her/him. As a result, you will receive a "bad" reputation

because of your uncooperative behavior and will not be helped by other *IS* players until you help someone else and restore your good reputation. For this reason, you are better off cooperating with anyone regardless of his/her reputation. Moreover, although the possibility that free-riders invade the population is completely removed, the problem still remains if there is even a small possibility of *errors in executing cooperation* (Boyd, 1989; Sugden, 1986). In reality, we are exposed to a risk of failure to help by accident. Being late for an appointment by oversleeping and passing by a dropped wallet or a donation box because of our own pressing business are a few examples. In an *IS* population, not only a free-rider but also just one error causes a chain of unfortunate defections ad infinitum, resulting in the breakdown of cooperative equilibrium. A core of this problem is that the *IS* cannot distinguish defection *on* free-riders from defection *by* free-riders.

(b) Second-order information

The chain of defection can be solved by *the standing strategy* (*ST*) that distinguishes two type of defectors: justified defectors and unjustified defectors (Leimar & Hammerstein, 2001; Panchanathan & Boyd, 2003). If you behave according to *IS*, you withhold help from a "bad" person, but you do not have any exploitative intent. This is the defection that *ST* sees as justifiable, and *ST* assigns this player a good standing. On the other hand, free-riders deny helping everyone. Therefore, if a player defects on someone in good standing, this behavior is regarded as unjustified defection and she/he will lower their standing to "bad." Accordingly, using *ST*, one needs to take into account not only its current partner's standing but also the current partner's

previous partner's standing (Figure 1). The latter is called *second-order information*, which is quite essential for indirect reciprocity to stabilize cooperation. In fact, Ohtsuki and Iwasa's (2004, 2006) series of exhaustive mathematic analyses revealed that only eight (out of 4096 possible) strategies, called the "leading eight," were able to stabilize cooperation. Although *ST* is one of the eight variants, it is important to emphasize that all of them make use of the second-order information to distinguish justified defectors from unjustified defectors.

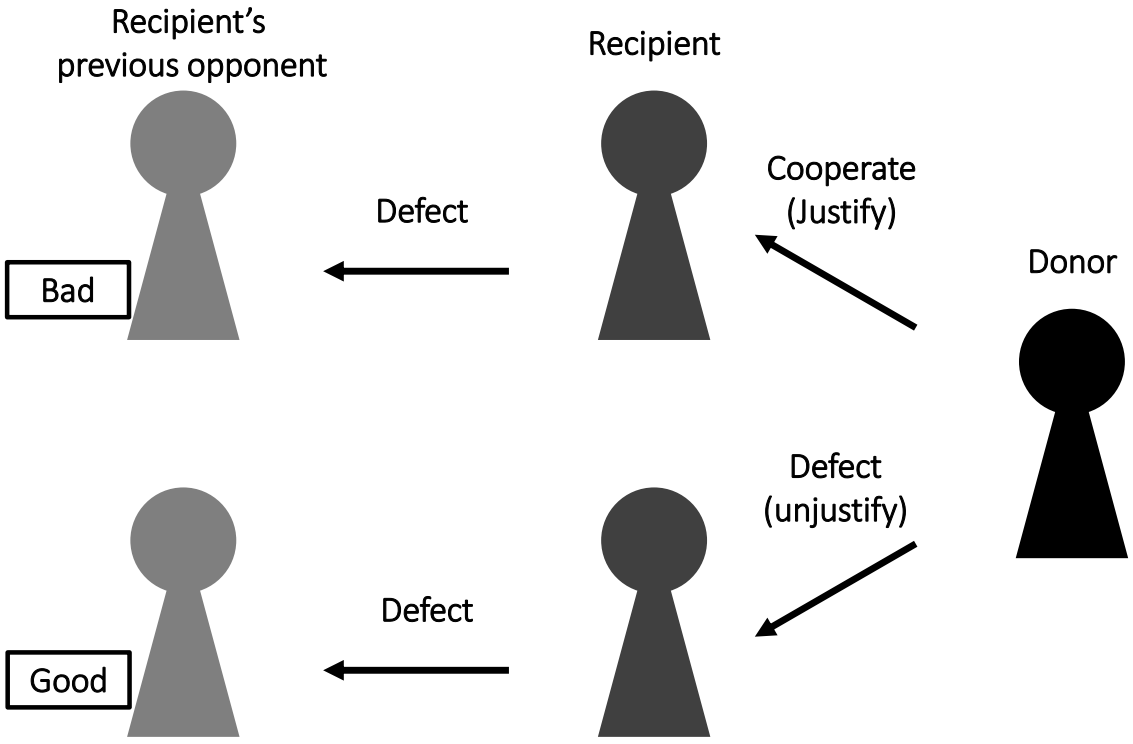

Figure 1. A schematic representation of the standing strategy (*ST*). The *ST* player (the donor) withholds help from the lower recipients who defected on a good player in the previous round. The *ST* player helps the upper recipient who defected on a bad player in the previous round.

Although *ST* is an evolvable strategy in theory, whether people behave in an *ST*-like manner is an empirical issue. If people use *ST*, it is predicted that they utilize second-order information in deciding whether to help someone. However, people who participated in experimental games actually did not robustly use *ST*. Although earlier studies reported negative results (Milinski, Semmann, Bakker, & Krambeck, 2001; Ule, Schram, Riedl, & Cason, 2009), there are some recent studies reporting positive results (Raihani & Bshary, 2015; Swakman, Molleman, Ule, & Egas, 2016).

Why is empirical evidence not substantially consistent with results of theoretical works? This could be because that second-order information is cognitively too demanding for people to utilize (Milinski et al., 2001). In fact, we utilize cognitive resources miserly on a daily basis (Fiske & Taylor, 2013). Moreover, it is known that people do not use all the relevant information to make rational decisions and tend to rely on relatively lower-order information in economic games (Ohtsubo & Rapoport, 2006). These empirical findings indicate that, even if reputation-assignment system is highly refined to achieve indirect reciprocity, a model employing much information for the refinement cannot be empirically valid. Therefore, increasing more information appears to be an unrealistic solution.

1.2. Reputation maintenance through the signaling of benign intention

Note that traditional models in the indirect reciprocity literature have implicitly assumed that actors are not involved in the process whereby their reputation is

determined. On the contrary, it has been empirically known that people attempt to *actively manage their impressions of themselves*. For example, in Milinski et al.'s (2001) experiment, justified defectors subsequently increased their cooperative behavior as if they communicated their lack of exploitative intent to other players. Likewise, other studies have shown that people who unintentionally treated their partners in an unfair manner engaged in apologizing and/or inflicting self-punishment (Ohtsubo & Watanabe, 2009; Tanaka, Yagi, Komiya, Mifune, & Ohtsubo, 2015; Watanabe & Ohtsubo, 2012). In these instances, although justified or unintentional defectors did not explicitly indicate their non-malicious intent, their behaviors implicitly (but reliably) indicate that they are not greedy exploiters. Therefore, when justified defectors want to communicate their non-malicious intent to recover their reputation, the use of these sorts of signals is possibly effective. If people produce such signals to communicate their intent in an indirect reciprocity context, they need not bother to use second-order information.

Based on the above argument, in this paper, I propose a new strategy for indirect reciprocity, *the intention signaling strategy* (*intSIG*). This strategy produces a signal when it defects on bad players and regards other signaling defectors as good player. In the subsequent chapters, I first introduce the details of this strategy and then present an evolutionary game analysis and a simulation study showing that the *intSIG* is theoretically robust. Thereafter, I report the results of two experiments revealing that people actually behave in an *intSIG*-like manner. Through these studies, I would like to shed light on an importance of social signals, in particular, how crucial role an active

intention signaling plays.

Chapter 2

Theoretical study

To examine the theoretical validity of *intSIG*, I conducted an evolutionary game analysis and computer simulation. Although they are explained more in detail in the following sections, I herein introduce them briefly.

The main purpose of the evolutionary game analysis is to seek a condition under which a free-rider cannot invade in a group consisting of *intSIG* players. If free-rider does so, indirect reciprocation through *intSIG* cannot evolve. Moreover, another condition under which unconditional cooperators cannot invade in the *intSIG* group is also examined. The reason is that those who cooperate with everyone allow free-riders to exploit them. Therefore, if unconditional cooperators can increase in the *intSIG* group, the sub-group of unconditional cooperators may allow the invasion of free-riders.

On the other hand, the purpose of the computer simulation is to examine whether the signal option can bring greater payoff for players than second-order information. As mentioned in Chapter 1, previous studies have suggested that indirect reciprocity is evolvable by using second-order information to distinguish justified defectors from unjustified defectors (Leimar & Hammerstein, 2001; Ohtsuki & Iwasa, 2004, 2006; Panchanathan & Boyd, 2003). Instead, the present study advocates for the superiority of using signal option as a plausible explanation for the evolution of indirect reciprocity. However, *intSIG* seems to be a less efficient strategy to achieve cooperative equilibrium because players pay a cost not only when they cooperate but also to produce a signal after

defection, whereas the *ST* players pay the cost only when they cooperate. It seems to indicate that possibility of an implementation error, which replaces players' cooperation with defection against their will, reducing *intSIG* players' payoff more than that of *ST* players. Therefore, I compare the net payoff of players in an indirect reciprocity context when they use the signal option with when they use second-order information. Altogether, the theoretical robustness of *intSIG* (it is evolutionarily stable against other major alternative strategies and can attain efficient cooperative equilibrium) is demonstrated.

2.1. *intSIG* in an indirect reciprocity context

To precisely define *intSIG*, I first explain the standard indirect reciprocity setting: there is an infinitely large population of individuals who engage in *a donation game*. This game consists of multiple rounds, and all players start the game with a good standing. In each round, players are randomly paired with one of the other individuals, then assigned either the role of a donor or recipient with the same probability, 0.5. To avoid direct reciprocation, the pairs of players never meet again. In the next step, donors decide whether they want to cooperate with the recipient or not. If they cooperate, they incur a cost (c) to confer a benefit (b) on recipients, and otherwise they save the cost without any earnings for the recipients ($b > c > 0$). However, there is a small probability ($e > 0$) of an implementation error whereby each donor fails to cooperate despite her/his intention to cooperate. Following the standard definition, erroneous cooperation (donors who intend to defect unintentionally cooperate) was not included in the implementation error. In addition to these settings, when donors decide whether to help, they are

informed of recipients' reputation. (In the first round, every player has good standing.) The reputation information is used by *IS* players, but not by unconditional players. *IS* players regard recipients who defected in a previous round as bad players and defect against them. On the other hand, *ALLC*, which always cooperates, and *ALLD*, which always defects, do not utilize any such information. After every donor has made her/his decision, the next round will occur with a probability of ω ($0 < \omega < 1$). Therefore, the expected number of the rounds in a game is $1 + \omega + \omega^2 + \ldots = 1/(1-\omega)$.

Based on these fundamental rules, *intSIG* can be described as follows. This strategy involves basically cooperating with others in good standing unless an implementation error occurs and defecting against others in bad standing. However, *intSIG* has another behavioral option. After an implementation error or intentional defection, donors subsequently produce *a costly signal*. If donors emitted the signal after defection, they can be regarded as good players by other *intSIG* players. In other words, this signal represents a lack of defectors' malicious intent. Note that the signal must be costly to inhibit free-riders from disguising themselves as cooperative players (cf. Grafen, 1990; Ohtsubo & Watanabe, 2009; Zahavi & Zahavi, 1997). If the signal cost (s) is cheaper than the cooperation cost (c), free-riders can fake the signal to maintain good standing. Therefore, the signal cost must be equal to or greater than that of cooperation to curtail the incentive to fake the signal.

Since *intSIG* players always produce the signal after defection and maintain good standing, they are never defected by other *intSIG* players except when partners commit an implementation error. However, if someone who uses some other non-

signaling strategy defects on a partner for whatever reason, she/he will be regarded as a bad player by *intSIG* players. In sum, in the group of *intSIG* players, regardless of the presence of implementation errors, all players maintain their good standing except some mutant players who use non-signaling strategies. Altogether, the payoffs of the two players in each round and the donor's standing in the next round determined by the *intSIG* are summarized in Table 1.

Table 1

The Payoff of the Donor and Recipient as a Function of the Donor's Behavior and Standing in the Next Round Determined by intSIG

Donor's Behavior	Donor	Recipient	**Donor's Standing in the Next Round**
Cooperation	$-c$	b	**good**
Defection with a Signal	$-s$	0	**good**
Defection without a Signal (only applied to mutant players)	0	0	**bad**

2.2. Evolutionary game analysis

(a) Evolutionary stability

Before describing how *intSIG* works in repeated interactions, I explain the evolutionary game analysis of *evolutionarily stability* (Maynard Smith, 1982; Maynard Smith & Price, 1973). This analysis examines the condition under which the focal strategy can prevent a rare alternative strategy from invading. Suppose that one invader

(Y) slips into a population of X. Since this population is assumed to be composed of an infinitely large number of focal strategies (X), all we have to do is to compare (i) X's payoff when playing with another X and (ii) Y's payoff when playing with X. If (i) is larger than (ii), Y will be eventually weeded out because its fitness is lower than X's fitness in this population. Therefore, it is concluded that *the focal strategy is an evolutionary stable strategy (ESS) against the alternative, invading strategy*. In standard ESS analyses for the evolution of cooperation, an *ALLD* (i.e., a free-rider), who defects against every other player, and an *ALLC*, who always cooperates regardless of others' standing, are typical invaders. This is because an *ALLD*'s invasion destabilizes a cooperative equilibrium, and the presence of a subgroup of *ALLC* can also destabilize the cooperative equilibrium by allowing the *ALLD* to invade the population of a mixture of the focal strategy and ALLC. Therefore, in this section, the evolutionary stability of *intSIG* against *ALLD* and *ALLC* was tested.

(b) *intSIG*'s payoff as a focal strategy

First, an *intSIG* player's payoff when playing with another *intSIG* player was computed. When all group members are *intSIG* players, one of the players in this group in each round earns $(1-e)(-c)+e(-s)$ as a donor (this player cooperates with the probability of $1-e$, while unintentionally fails to do so and produces the costly signal with the probability of e). Because of the costly signal after implementation errors, each *intSIG* player's standing is always good. Therefore, the *intSIG* player earns $(1-e)(b)$ as a recipient in each round. As the donor and recipient roles are assigned with the same

probability, their expected payoff in each round, w_{SIG} can be written as:

$$w_{SIG} = \frac{(1-e)(b-c)-es}{2} \quad . \tag{1}$$

Since this game continues with the probability of ω, the net payoff of *intSIG* players is W_{SIG}:

$$W_{SIG} = \frac{1}{1-\omega}\frac{(1-e)(b-c)-es}{2} \ . \tag{2}$$

(c) *intSIG*'s evolutionary stability against *ALLD*

In this section, I examined the condition under which *intSIG* can be stable against *ALLD*. First, an *ALLD* player's expected payoff when playing with an *intSIG* player was calculated. Since it is assumed that the frequency of *ALLD* is negligible, the net payoff of *intSIG* players is written as Eq. (2).

When an *ALLD* player is a donor, it pays 0 because of withhold cooperation toward the recipient. When this player is a recipient, she/he earns either $(1-e)b$ when her/his standing is good or 0 when it is bad. Let $G_{ALLD}(t)$ be the probability that the *ALLD* player is in good standing after t-th round. Since it is assumed that all players are in good standing when the game starts, $G_{ALLD}(0) = 1$. The *ALLD* player's standing falls into bad once assigned to the donor role and never returns to good. Accordingly, an *ALLD* player in good standing will shift to bad standing with the probability of 0.5, which is the probability that it will be assigned to the donor role.

$G_{ALLD}(t+1) = G_{ALLD}(t)\times(1/2)$.

Therefore,

$$G_{ALLD}(t) = \left(\frac{1}{2}\right)^t. \quad (5)$$

The *ALLD*'s payoff in the *t*-th round, $w_{ALLD}(t)$, is calculated by taking account of the probability of being in good standing, the probability of being assigned to the recipient role, and the benefit conferred by a cooperative donor (the payoff when an *ALLD* player is assigned to the donor role is always 0):

$$w_{ALLD}(t) = \left(\frac{1}{2}\right)^{t-1}\frac{1}{2}(1-e)b = \left(\frac{1}{2}\right)^t(1-e)b. \quad (6)$$

Since the next round will occur with the probability of ω, the net payoff of the *ALLD* is:

$$W_{ALLD} = \left(\frac{1}{2}\right)(1-e)b + \omega\left(\frac{1}{2}\right)^2(1-e)b + \omega^2\left(\frac{1}{2}\right)^3(1-e)b + \cdots$$

$$= \frac{1}{2-\omega}(1-e)b. \quad (7)$$

Based on Eq. (2) and Eq. (7), an *ALLD* player cannot invade a group of *intSIG* players as far as the following condition holds:

$W_{SIG} > W_{ALLD}$

$$\Leftrightarrow \frac{1}{1-\omega}\frac{(1-e)(b-c)-es}{2} > \frac{1}{2-\omega}(1-e)b$$

$$\Leftrightarrow \frac{1-e}{e}\{(2-\omega)(b-c) - 2(1-\omega)b\} > s(2-\omega). \quad (8)$$

In the Eq. (8), if it is assumed that the error rate (e) is small, $\frac{1-e}{e}$ takes a large positive value. Furthermore, the right side of the inequality is always positive (both s and $2-\omega$ take positive values by definition). Therefore, Inequality (8) holds if $(2-\omega)(b-c) - 2(1-\omega)b > 0$. This condition can be reduced as follows:

$$\omega > \frac{2c}{b+c} \quad . \quad (9)$$

As $b > c$, the range of the right side of Inequality (9) is $0 < \frac{2c}{b+c} < 1$, which corresponds to the range of ω. Therefore, Inequality (9) reveals that, when the implementation error rate e is small, *intSIG* is stable against *ALLD* as far as the game continues with a probability greater than $\frac{2c}{b+c}$. For example, when $b = 2$ and $c = 1$, condition (9) only requires that the games has to consist of more than 3 rounds on average (i.e., $\omega > 2/3$). It is noteworthy that this condition does not depend on the size of the signal cost, s.

Remember that the signal cost, s, needs to be equal to or greater than the cost of cooperation, c. Substituting c for s in Inequality (8) yields the following condition:

$$e < 1 - \frac{(2-\omega)c}{\omega b} \quad . \tag{10}$$

This condition holds when the game continues substantially long. For example, when ω is nearly 1, this condition becomes $e < 1 - \frac{c}{b}$. Therefore, if the game continues substantially long and e is sufficiently small, *ALLD* cannot invade the group of *intSIG* players.

(d) *intSIG*'s evolutionary stability against *ALLC*

I next explored the condition under which *intSIG* is stable against the invasion of *ALLC*. When there is no possibility of implementation errors, rare *ALLC* players and *intSIG* players will peacefully co-exist in cooperative equilibrium. However, if the possibility of implementation errors is introduced, the payoffs of the *intSIG* and *ALLD* will diverge because *intSIG* players can maintain their good standing by producing a costly signal, while *ALLC* players have to wait one donor-round to cooperate and restore

their good standing.

To obtain the net payoff of *ALLC* in the *intSIG* group, let $G_{ALLC}(t)$ be the probability that *ALLC* is in good standing after the *t*-th round. We have $G_{ALLC}(0) = 1$ as an initial condition. The *ALLC* player's standing becomes bad only when committing an implementation error. Therefore, after playing the donor role, her/his standing is good with the probability of $1-e$. After playing the recipient role, her/his standing does not change. Accordingly, the probability that the *ALLC* player is in good standing after the (t+1)-th round is

$$G_{ALLC}(t+1) = \frac{1}{2}G_{ALLC}(t) + \frac{1-e}{2} \quad . \tag{11}$$

Subtracting $1-e$ from the both sides of Eq. (11) yields

$$G_{ALLC}(t+1) - (1-e) = \frac{1}{2}G_{ALLC}(t) - \frac{1-e}{2} \quad . \tag{12}$$

Let $H_{ALLC}(t) = G_{ALLC}(t)-(1-e)$, and Eq. (12) can be rewritten as

$$H_{ALLC}(t+1) = \frac{1}{2}H_{ALLC}(t) \quad . \tag{13}$$

Notice that $H_{ALLC}(0) = 1-(1-e) = e$. Therefore,

$$H_{ALLC}(t) = G_{ALLC}(t) - (1-e) = e\left(\frac{1}{2}\right)^t \quad . \tag{14}$$

From Eq. (14), we obtained the probability that the *ALLC* player is in good standing after the *t*-th round as follows:

$$G_{ALLC}(t) = e\left(\frac{1}{2}\right)^t + (1-e) \quad . \tag{15}$$

Using Eq. (15), the expected payoff of the *ALLC* at the *t*-th round can be computed. If the *ALLC* plays the donor role, its payoff is $(1-e)(-c)$ regardless of its

standing. If the *ALLC* plays the recipient role, its expected payoff is (1−*e*) *b* when in good standing, while the expected payoff is 0 if it is bad. Accordingly, the *ALLC*'s expected payoff at the *t*-th round is written as

$$w_{ALLC}(t) = -\frac{1}{2}(1-e)c + \frac{1}{2}b(1-e)G_{ALLC}(t-1)$$

$$= \left(\frac{1}{2}\right)^t e(1-e)b + \frac{1}{2}\{(1-e)^2 b - (1-e)c\}. \quad (16)$$

From Eq. (16), the *ALLC*'s net payoff is derived as follows:

$$W_{ALLC} = \frac{1}{2-\omega}e(1-e)b + \frac{1}{1-\omega}\frac{(1-e)^2 b-(1-e)c}{2} \ . \quad (17)$$

Based on Eq. (2) and Eq. (17), the condition under which *intSIG* is stable against an *ALLC* ($W_{SIG} > W_{ALLC}$) is derived as follows:

$$\frac{1}{1-\omega}\frac{(1-e)(b-c)-es}{2} > \frac{1}{2-\omega}e(1-e)b + \frac{1}{1-\omega}\frac{(1-e)^2 b-(1-e)c}{2} \ ,$$

which is rewritten as

$$(2-\omega)e(1-e)b - (2-\omega)es > 2(1-\omega)e(1-e)b \ . \quad (18)$$

By dividing the both sides of Inequality (18) by $e>0$, the ESS condition of *intSIG* against an *ALLC* was further rewritten as below:

$$e < 1 - \frac{(2-\omega)s}{\omega b} \ . \quad (19)$$

Because I divided both sides of inequality by a small number, *e*, to obtain the condition (19), the difference between the net payoffs of *intSIG* and the *ALLC* is small. However, if condition (19) holds, *intSIG* is stable against the *ALLC*. This tends to hold when the cost of signal, *s*, is relatively small compared to the benefit of being helped, *b*. In other words, unlike the ESS condition against an *ALLD*, which did not depend on the cost of

the signal, *intSIG* is less likely to be stable against an *ALLC* if the signal cost is large.

I further examined condition (19) assuming that the signal cost, s, is equal to the cost of cooperation, c. Interestingly, the resultant condition was exactly equal to the condition under which *intSIG* was stable against the invasion of the *ALLD*, which is condition (10)

$$e < 1 - \frac{(2-\omega)c}{\omega b} \quad . \tag{20}$$

(e) Summary of mathematical analyses

I investigated under which condition *intSIG* is evolutionarily stable against an *ALLD* and *ALLC*. First, *intSIG* was stable against an *ALLD* as far as the interactions continue sufficiently long and the stability condition does not depend on the cost of the signal. Second, although the *intSIG* and *ALLC* players' expected payoffs were close to each other, *intSIG* was stable against the *ALLC* when the cost of the signal was not too large. When it is assumed that the cost of the signal, s, was equal to the cost of cooperation, c, which is a sufficient amount of signaling cost to prevent dishonest signalers from undermining the separating equilibrium, it was shown that *intSIG* was stable against both the *ALLD* and *ALLC* under exactly the same condition. Therefore, it can be concluded that the group of *intSIG* players is evolutionarily stable.

2.3. Computer simulation

Additional to the analysis of ESS, I examine whether *intSIG* is more efficient

to achieve a high level of net payoff than *ST*. *IntSIG* introduces a signal option into players' behavioral option to solve an enduring problem of indirect reciprocity, which is how to distinguish justified defectors from unjustified defectors. On the other hand, *ST* has been proposed by previous studies to solve the same problem by utilizing the second-order information to detect the defectors' type. Due to the costliness of the signal, *intSIG* appears to be less efficient than *ST*, and if actually so, theoretical validity of *intSIG* relative to that of *ST* is tarnished to some extent. Accordingly, I conducted a computer simulation to directly compare the net payoffs between two groups containing either *intSIG* players or *ST* players.

(a) Method

Similar to the evolutionary game analysis, the donation game was employed to compute the net payoff of the groups. The specific settings of the game were as follows: There were two groups which consisted of 20 players. One of these groups comprised 20 *intSIG* players and another comprised 20 *ST* players; hence, each group was not a mixture of players using different strategies. The cost of cooperation (c) was a constant of integral as one, while the benefit of receiving cooperation (b) was a variable ranging from 1.1 to 4 ($b > c = 1$). The probability of an implementation error (e) was a variable ranging from 0 to 0.1. Accordingly, I ran a set of simulations of the donation game under 7 (b = 1.1, 1.5, 2, 2.5, 3, 3.5, or 4) $\times$ 4 (e = 0, 0.01, 0.05, or 0.1) conditions. There were 100 rounds for each donation game. At the end of the one game, a sum of the net payoffs of each player in a group was computed. This process was repeatedly

simulated 10000 times, and the sum of the net payoff was averaged.

As for the *intSIG* group, when donors fail to cooperate due to an implementation error, they immediately produce a signal with a cost. The amount of the signal cost (s) was set to be equal to the cooperation cost ($s = c = 1$). Recall that, according to the behavioral and reputation-assignment strategies used by *intSIG*, players acquire bad standing only when they defect without producing a signal. Therefore, in a group of *intSIG* players, no one's standing falls into bad, and any defection in this group is due to an implementation error.

On the other hand, according to the behavioral and reputation-assignment strategies that *ST* uses, players acquire bad standing when they defect on recipients in good standing, while players maintain their good standing when they defect on recipients in bad standing. Cooperation is always regarded as good behavior. Due to an implementation error, *ST* players sometimes fail to cooperate with recipients in good standing, and consequently become bad players. Of course, those who defect on the recipient with bad standing due to the previous implementation error are considered justified defectors, and their standing will not become bad.

(b) Result

It is noteworthy that if an implementation error never occurs ($e = 0$), the average net payoffs of an *intSIG* group and a *ST* group are equal because players in both groups always cooperate throughout the 100 rounds. Therefore, I set the condition when $e = 0$ was a standard, then computed the relative amounts of the net payoff as a function

of *e* and *b*.

The results are shown in Figure 2. In both groups, as the error rate, *e,* became larger, the relative net payoff went down. For the ST group, however, when *e* was fixed, the relative payoff was not changed regardless of *b* (purple bars). On the other hand, the relative payoff of the *intSIG* group was influenced both by *e* and *b* (blue bars). If *b* was small, the relative payoff of the *intSIG* group was quite less than that of the ST group. However, when *b* was larger than 2, the relative payoff of *intSIG* outweighed that of the *ST* group regardless of *e*.

These results indicated that if the benefit-to-cost ratio is sufficiently large, *intSIG* is a more efficient strategy to achieve high cooperative equilibrium than *ST*. Interestingly, this difference in efficiency does not emerge until a chance of an implementation error exists, which is consistent with real life. However, why did the difference emerge? This is possibly because when *ST* players fail to cooperate, they cannot receive benefits from other *ST* players until they cooperate the next time. On the other hand, when *intSIG* players fail to cooperate, they immediately pay a cost to produce a signal and they can obtain benefit from other *intSIG* players. It means that the potential loss of *ST* players who fail to cooperate depends on the amount of benefit of cooperating, while that of *intSIG* players depends on the amount of the signal cost, which is assumed to be equal to the cooperation cost. Therefore, the more the benefit-to-cost ratio of cooperation increases, the larger the difference in the potential loss between *ST* and *intSIG* players becomes. The simulation results clearly showed that the signal option can preserve the chance of benefitting from others' cooperation more efficiently than using second-order

information.

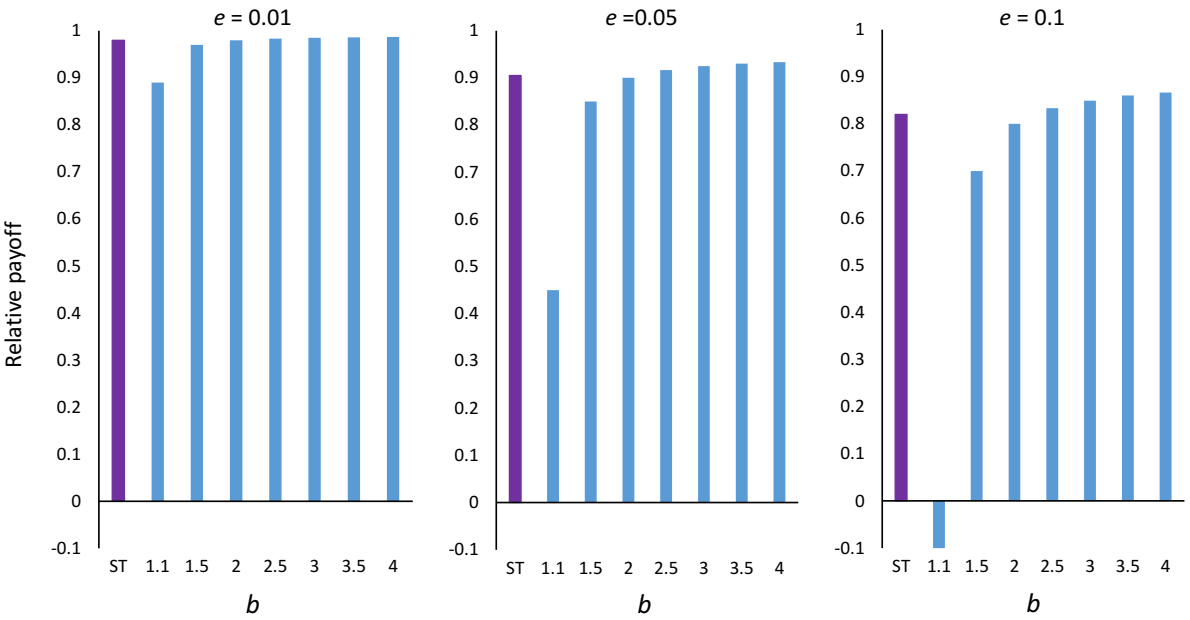

Figure 2. The relative net payoff of the *intSIG* group and the *ST* group as a function of error rate (*e*) and the relative amount of benefit (*b*) to the cost when the net payoff under the condition where $e = 0$ was set to be a standard (100 %). Since the payoffs of the *ST* group was not changed depending on *b*, they are shown on the left side of the graphs (the purple bars). All blue bars indicate the payoffs of the *intSIG* group.

Chapter 3

Empirical study

In this chapter, I introduce two experiments examining whether people actually behave in an *intSIG*-like manner in an indirect reciprocity context. Specifically, the purpose of the empirical study was to test whether people send signals when they defect without malicious intent, and whether to receive other players' signal as an absence of this intent.

As mentioned earlier, *intSIG* uses the signal option to distinguish justified defectors from unjustified defectors. This option enables *intSIG* players to solve the problem of the chain of defection with which *image scoring strategy* (*IS*) has faced. Previous studies have proposed the *standing strategy* (*ST*) to conquer this problem. *ST* players attend to not only their current partner's behavior (first-order information) but also their current partner's previous partner's reputation (second-order information). Although theoretical works have shown that *ST* can maintain a cooperative equilibrium in the presence of implementation errors and a few uncooperative players, empirical evidence examining whether people actually utilize second-order information is mixed. On the other hand, the theoretical validity of *intSIG* was supported in the studies described in Chapter 2, while its empirical validity remains to be demonstrated (Table 2). In addition to the verification of *intSIG*'s empirical validity, I herein also examined whether people behave in a *ST*-like manner to retest the results of related studies (Milinski et al., 2001; Swakman et al., 2016). Accordingly, participants played the donation game (Wedekind &

Milinski, 2000) either in the signaling condition or the standing condition.

Table 2

The Properties and the Validity of the Standing strategy and the Intention Signaling Strategy

	Standing	Intention Signaling
second-order information	○	×
signaling option	×	○
theoretical validity	○	○
empirical validity	mixed	?

Before explaining the details of the two experiments, I explain the rule of the donation game, which was common to both experiments. As explained in Chapter 2, each round of the donation game consisted of two phases: random pairing and donors' decision. At first, participants were randomly paired with another putative participant and randomly assigned to the role of either the donor or the recipient. In fact, participants interacted with pre-programed computerized players instead of other participants because real interaction may prevent participants from facing every possible type of reputation information, which is described later. After the pairing, donors decided whether to give the benefit (b) to the recipient while paying the cost (c). Besides, in the signaling condition, participants who denied giving b for whatever reason subsequently decided whether to

abandon *c* (= to use signal option), which was once kept by denying to give *b*. In consequence, when donors decided to give, they could see the recipients' previous behavior (i.e., reputation information) as three types, "gave," "did not give + abandoned," and "did not give + did not abandon." On the other hand, in the standing condition, donors were presented with second-order information about the past behavior of their recipients' previous partner. Therefore, donors were exposed to one of the four types of information that can be expressed by a combination of first-order information and second-order information (Table 3). Since participants in each condition faced every type of reputation information throughout the game, the strategies that each participant utilized can be estimated.

Moreover, donors who decided to give *b* were exposed to a risk that they would fail to do it accidentally. If this implementation error occurred, their reputation information became "did not give" as long as they immediately abandoned *c*. This indicates that participants who were engaged in the standing condition did not have a means to modify their own reputation information after the error except giving *b* in the following rounds.

Table 3

Reputation Information with which Donors Were Presented in the Donation Game of the Signaling and Standing Conditions

	the signaling condition	the standing condition
first-order information	gave, did not give + abandoned, did not give + did not abandoned	gave, did not give
second-order information	×	had given, had not given

3.1. Hypotheses

In the signaling condition, I tested three hypotheses. Hypotheses 1a and 1b are about signalers' behaviors. Hypothesis 2 is about signal receivers' reaction to the signal. Hypothesis 1a is based on the operationalization of three types of defection: an "implementation error," a "justified defection," and an "unjustified defection." An "implementation error" is a defection replaced with one's given choice with a computer-programed probability. A "justified defection" is a refusal to give *b* to recipients who are in bad standing. An "unjustified defection" is the refusal to give *b* to recipients who are in good standing. Ideally, *intSIG* players always abandon *c* after an implementation error and justified defection, while they never commit an unjustified defection. Therefore, the first prediction is as follows:

Hypothesis 1a: Participants abandon *c* more frequently after implementation errors and justified defections than unjustified defections.

The above classification of defection also allows us to categorize defectors into

two types: “Justified defectors” and “Unjustified defectors”. Justified defectors are discriminators who give *b* to the recipients in good standing but not to the recipients in bad standing. Unjustified defectors are free-riders who refuse to give *b* to the recipients regardless of their standing. If justified defectors want to protest that they are different from unjustified defectors to maintain their standing, they should abandon *c* because of the absence of second-order information.

Hypothesis 1b: Justified defectors abandon *c* more frequently than unjustified defectors.

Even if participants actually use the signal option, cooperative equilibrium cannot be achieved unless participants react to others abandoning behavior to distinguish justified defection from unjustified defection. Since abandoning is linked to be justified, the following hypothesis is derived:

Hypothesis 2: Participants give *b* more frequently to other players who gave *b* (“gave”) or refused to give it but abandoned *c* (“did not give but signaled”) than to players who refused both to give *b* and to abandon *c* (“did not give”) in the previous round.

In the standing condition, donors were faced with recipients with four types of reputation information: GG, GN, NG, and NN. Each symbol indicates the combination of first-order information (the left side: G or N) and second-order information (the right side: G or N), where G represents “gave” and N represents “did not give.” For example, the sign GN means that, in the previous round, the current recipient gave *b* to her/his partner who had not given *b*.

If participants distinguish justified defections from unjustified defections based

on second-order information, they may regard recipients with information NN as justified defectors and recipients with NG as unjustified defectors. Therefore, it can be hypothesized that participants give b more frequently to other players whose reputation information are GG, GN, and NN than to players whose reputation information is NG. On the other hand, if participants do not use second-order information, it is hypothesized that participants give *b* more frequently to other players with information GG and GN than to players with information NG and NN.

3.2. Experiment 1

The matter I have argued thus far is the evolution of human cooperation. It assumes that participants would have a tendency to behave in an *intSIG*-like manner *prior to the experience*. Accordingly, in experiment 1, participants were not informed of how their behavior was responded to by other players. The reason is that if they believe they will receive a benefit from other players after abandoning *c*, they may learn that using the signal option is effective to maintain good standing. Therefore, the specific purpose of experiment 1 is to test the hypotheses described above *without the possibility of social learning*.

(a) Method

Participants. One hundred seven undergraduates (62 males and 45 females) at Kobe University participated in the experiment. The mean $\pm$ *SD* of participants' age was 19.17 $\pm$1.09. To recruit participants, they were informed that 500 JPY (a show-up fee) and

extra monetary earnings depending on the outcome of experimental game would be paid for their participation. In fact, participants were received 1000 JPY as the extra, which was beyond the maximum amount they could earn through the game. Three participants were dropped from the analyses, as they casted doubt on the presence of other players and requested to remove their behavioral data after debriefing.

Procedure. Experiment 1 was composed of the donation game and post-game questionnaire. At first, all participants were guided into a laboratory with six separated cubicles where they could not see potential other players, entered one of the cubicles, and signed the informed consent form. There was a computer in each stall and participants played the donation game with it. To prevent the possibility that direct reciprocation would arise, participants were informed that the game would be played anonymously.

Before playing either the signaling condition or the standing condition, participants first engaged in the practice session, which was the donation game including neither the signal option nor second-order information. This game consisted of 50 rounds and participants were not informed prior to remove effects of the shadow of the future. Participants were instructed that they would play the game with 5 other participants, but they actually interacted with a preset computer program. In each round, participants were randomly assigned to either the donor or recipient role. They played 25 rounds of the game as a donor, and 25 rounds of the game as a recipient. However, they inevitably played as the recipient role in the first round, as the recipients in this round do not have any behavioral histories. In other words, this setting was intended to

prevent participants from reacting to a recipient with no reputation information.

When participants were assigned to the donor role, they received an endowment of 5 JPY and decided whether to "give" it to their recipient or "not give" it and retain it for themselves. If participants chose "give," they would lose 5 JPY and the recipient would receive 10 JPY, which was double the donors' resources ($b > c$). If they chose "not give," they would keep 5 JPY and the recipient would receive nothing. However, their "give" choices were changed to "not give" with the probability of .10. Participants themselves immediately knew when this implementation error occurred, whereas their current recipient and prospective partners in the following rounds could not know whether they made' "not give" choices intentionally or due to the error. Participants understood the presence of the error, but not the specific probability of it. Throughout the session, all participants interacted with recipients with "gave" or "did not give" histories approximately 13 and 12 times, respectively, so that they could use only first-order information. After determining their behavior, participants' cumulative acquired money, which was displayed at the bottom of computer display, was upgraded.

When participants were assigned to the recipient role, all they had to do was wait for (pseudo) the current donor's decision. After a few seconds, the pre-programed duration ranging from 3 to 10 seconds, the next round started. Note that whether participants received 10 JPY from the current donor had to remain unclear to them, as this information might allow them to make inferences about what kind of strategies the other players were employing. For example, if they received 10 JPY after they had chosen "give" and did not receive 10 JPY after they had chosen "not give," they might

infer that other players make their "give" choice contingently on their previous choice. Therefore, participants only received an aggregated form of feedback after playing the recipient role five times. In other words, after every five rounds when participants were assigned to the recipient role, they were informed of cumulative earnings ranging from 0 (receiving 10 JPY from no donors) to 50 JPY (receiving 10 JPY from every donor). Moreover, the amount of this feedback was randomly determined. Hereby, participants could not infer the *IS*-like association between their behavior as the donor role and whether they received money as the recipient role. (i.e., If I choose "give" in the current round, I will receive 10 JPY from someone else.) When participants received the feedback, their cumulative earning on the display was updated.

After the practice session, 53 and 54 participants engaged in the signaling condition and standing condition, respectively. Participants were instructed on the additional rules corresponding to their condition. However, again, they were not informed of the exact number of the rounds (100 rounds) to remove effects of the show of the future. As with the practice session, the total rounds participants were assigned to a donor role and a recipient role were equal (50 rounds, respectively) but to which roles they were assigned in each round was randomly determined.

In the signaling condition, participants as donors could use the signal option after "not give" regardless of whether the choice was made intentionally or due to the implementation error. In particular, participants were asked whether they would like to "abandon" or "not abandon" the 5 JPY that they kept by "not give." If participants chose "abandon," they would earn nothing in the round. If they chose "not abandon,"

they would keep 5 JPY. Therefore, donors were exposed to three patterns of behavioral histories, “gave,” “did not give + abandoned,” and “did not give + did not abandon.” In the following rounds, they would be assigned to the recipient role. To prevent this abandonment from explicitly indicating the signaling behavior, the term “signal” was not used in the condition. Participants interacted with recipients with each history approximately 25 (“gave”), 13 (“did not give + abandoned”), and 12 times (“did not give + did not abandon”). Donors’ choice of whether to abandon 5 JPY or not did not affect current recipients’ benefit so that they would receive nothing regardless of their current donors chose “abandon” or “not abandon.”

In the standing condition, participants as donors were provided the information not only about whether a current recipient chose “give” or “not give” (first-order information), but also about whether the current recipient’s previous recipient, with whom a current recipient had interacted when she/he had been assigned to a donor role, had chosen “give” or “not give” (second-order information). This means that donors were faced with 2 (current recipient’s “gave” or “did not give” histories) × 2 (current recipient’s previous recipient’s “had given” or “had not given” histories) patterns of information. Although participants were presented with such information, as mentioned above, in this paper, these 4 histories are symbolized as GG, GN, NG, and NN for the sake of expedience. The left side of G/N represents “gave”/“did not give” and the right side of G/N represents “had given”/“had not given.” Participants interacted with recipients with each history approximately 13 (GG), 13 (GN), 12 (NG), and 12 (NN) times, respectively. Unlike the practice session and the signaling condition, participants

inevitably played the recipient role in the second round as well as in the first round, as second-order information was not available in the first two rounds. Note that both the signaling condition and the standing condition had an ambiguous feedback setting, which was same as the practice session; hence, participants could not infer the *intSIG*-like or *ST*-like association between their behavior as the donor and whether they received money as the recipient.

After they completed the experimental game, participants were asked to fill out a post-game questionnaire about the strategy they used in the game. In the questionnaire, participants presented recipients with all possible patterns of histories in the given condition and were asked to indicate (a) their impression of the recipient; (b) their inference of goodness of recipient's intention; (c) their hypothetical behavior ("give" or "not give") to the recipient. Participants rated their impression and the inferred goodness of recipients' intention on a five-point scale (1 = "very bad" to 5 = "very good") and stated how they would behave toward the recipient (either "give" or "not give"). In addition, participants engaged in the signaling condition were asked to indicate their hypothetical decision about the signal option. Participants who chose "give" further noted whether they would abandon 5 JPY as a reaction to an implementation error, and participants who chose "not give" simply stated whether they would abandon 5 JPY (either "abandon" or "not abandon"). After the post-game questionnaire, participants were told the purpose and hypotheses of the experiment, informed of the presence of deception about pseudo other players and its necessity, and given a right to remove their data from analysis. Regardless of their withdrawal to

provide their data, all participants received 1500 JPY.

(b) Results

Cooperation rate in the practice session. Before testing the hypotheses about *intSIG* and *ST*, I examined whether participants behaved an *IS*-like manner in the practice session. The reason is that, if their tendencies to choose "give" did not differ toward recipients with a history of "gave" versus "did not give," participants might not discriminate among recipients depending even on their binary histories. Then, the mean cooperation rates toward the recipients with each history ("gave" or "did not give") were computed separately. The result is that participants' mean cooperation rate was .71 (SD = 0.30) when the recipients had a "gave" history and .45 (SD = 0.26) when they had a "did not give" history. These data were submitted to a 2 (recipients' history: "gave" vs. "did not give") × 2 (game type: signaling vs. standing) ANOVA including the former factor was repeated measure. The analysis showed that the main effect of recipients' history was significant, $F_{1, 102} = 78.07$, $p < .001$, and other effects were not significant. These results indicate that participants were more likely to choose "give" toward recipients with a "gave" history than recipients with a "did not give" history. More importantly, the non-significant effect of the condition implies that the random assignment was successful (there was no systematic difference in their cooperativeness before receiving any instructions relevant to the experimental conditions).

Testing hypotheses about *intSIG*. Then, I examined whether participants used the signal option as the *intSIG* strategy does. I first tested Hypothesis 1a: Participants

abandon *c* more frequently after an implementation error and a justified defection than an unjustified defection. Recall that a justified defection means choosing "not give" toward recipients in bad standing, and an unjustified defection means choosing "not give" toward recipients in good standing. On the basis of the rule of *intSIG* strategy, we operationalized players with a "did not give + did not abandon" history as bad players, and players with "gave" or "did not give + abandoned" histories as good players. To compute the signaling rate, the numbers of signal options participants used were added up from each defection type, and then these total numbers were divided by the sum of each defection type in which participants were involved. As a result, the signaling rates as a function of defection types were .57 (implementation error), .23 (justified defection), and .07 (unjustified defection), respectively (Figure 3). Fisher's exact test with the Bonferroni correction revealed that differences of signaling rates between all possible pairs were significant (every $p < .017$). This means that participants used the signal option more frequently after implementation errors and justified defections than unjustified defections, and hence Hypothesis 1a was supported.

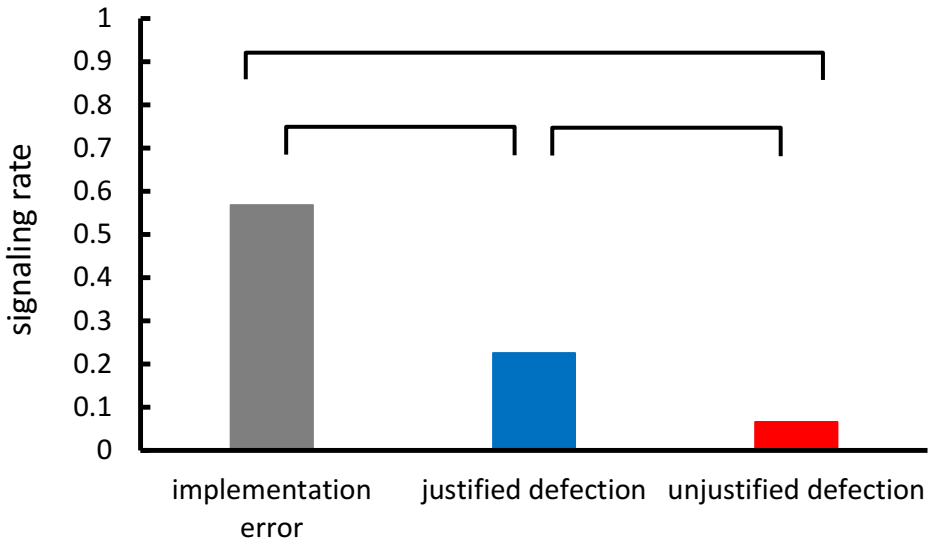

Figure 3. Signaling rate as a function of defection types in the signaling condition of experiment 1.

Second, I tested Hypothesis 1b: Justified defectors abandon *c* more frequently than unjustified defectors. Operationally, unjustified defectors were defined as participants who chose "not give" more than 80% of the time when they were paired with recipients in good standing (attached to "gave" or "did not give + abandoned" histories). Justified defectors were defined as those who were not categorized as unjustified defectors and chose "not give" more than 80% of the time when faced with recipients in bad standing (attached to "did not give + did not abandon" histories). According to these definitions, out of the 52 participants in the signaling condition, there were 19 justified defectors whose mean signaling rate was 0.341 (*SD* = 0.101) and 13 unjustified defectors whose mean signaling rate was 0.007 (*SD* = 0.000). As predicted, justified defectors used the signal option significantly more often than

unjustified defectors ($t_{30} = 3.77$, $p < 0.001$, two-tailed test).

Finally, Hypothesis 2 was tested: Participants gave *b* more frequently to other players who gave *b* or refused to give it but abandoned *c* than to players who refused both to give *b* and to abandon *c* in the previous round. As with the practice session, the mean cooperation rates toward the recipients with each history ("gave," "did not give + abandoned," or "did not give + did not abandon") were computed separately. As shown in Figure 4, the main effect of recipients' history was significant ($F_{2,102} = 34.83$, $p < 0.001$), and then a post hoc test using Ryan's method was employed. The results showed that participants chose "gave" toward recipients with a "gave" history more often than recipients with a "did not give + abandoned" history. More importantly, participants chose "gave" toward recipients with "did not give + abandoned" more often than recipients with "did not give + did not abandon," and thus Hypothesis 2 was supported.

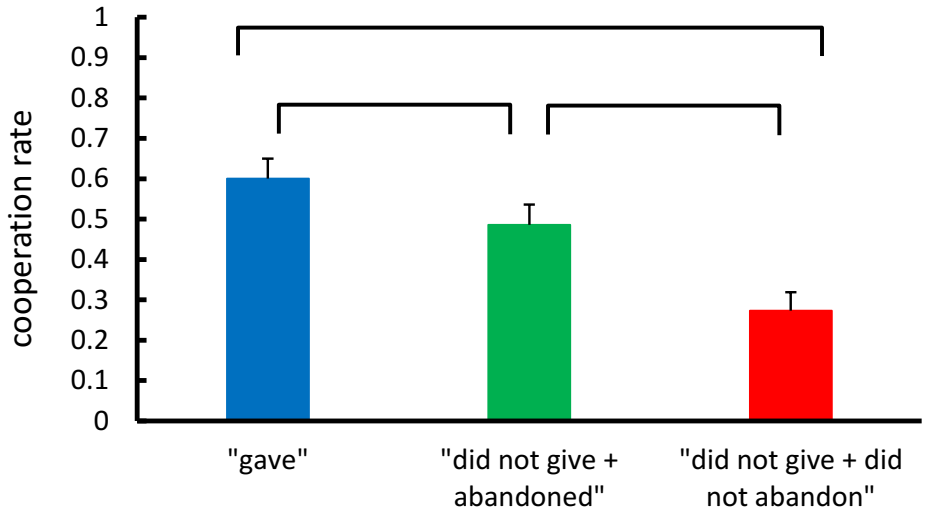

Figure 4. Cooperation rate as a function of recipients' histories in the signaling condition of experiment 1. The error bars indicate standard errors of the mean.

Testing hypotheses about *ST*. For the standing condition, a hypothesis that participants behave in an *ST*-like manner would be supported if they distinguish recipients with an NN history (i.e., justified defectors) from recipients with an NG history (i.e., unjustified defectors). The main effect of recipients' history was significant ($F_{3,153} = 20.49$, $p < 0.001$). However, a post hoc test using Ryan's method revealed that the mean cooperation rates toward recipients with an NN history (i.e., justified defectors) and recipients with an NG history were not significant (Figure 5). This result indicated that participants did not chose "give" toward justified defectors more than unjustified defectors. In other words, contrary to the prediction from the ST strategy, but consistent with the prediction from the *IS* strategy, participants did not treat the justified defectors more favorably than unjustified defectors.

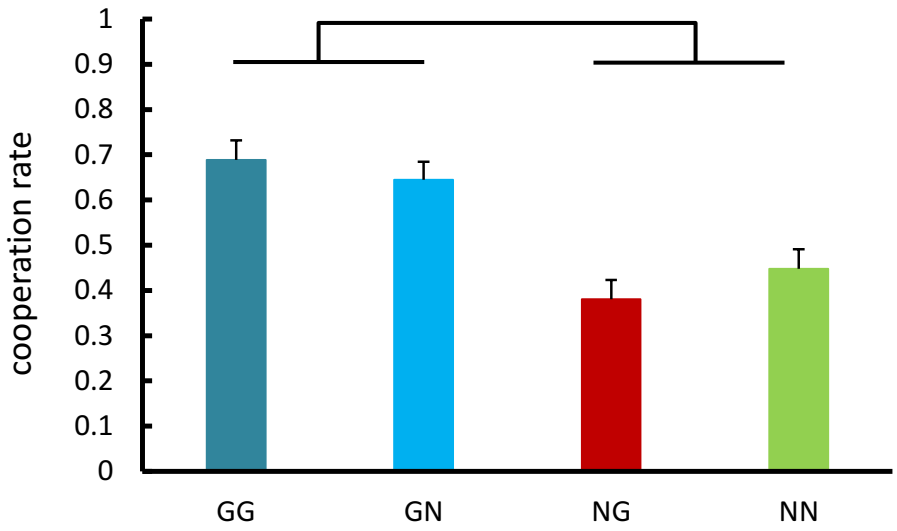

Figure 5. Cooperation rate as a function of recipients' histories in the standing condition of experiment 1. The error bars indicate standard errors of the mean.

Reaction time to recipients' history. The above results showed that participants were likely to use *intSIG* in an indirect reciprocity context but not likely to use *ST*. Milinski et al.'s (2001) study comparing the empirical validity between *IS* and *ST* suggested that second-order information is too cognitively taxing for people to use intuitively. If this cognitive load determines whether people use a certain strategy, then participants in the signaling condition would take less time to make their choice of "give" or "not give" than those in the standing condition. To test this prediction, the reaction time (RT) in the two conditions were compared. The mean RT was 2.61 seconds (SD = 1.17) in the signaling condition and 2.74 seconds (SD = 0.87) in the standing condition. Although this pattern followed the prediction, the difference between these two RTs was not significant (t_{102} = 0.61, p = .54, two-tailed test).

Post-game questionnaire. The results of the post-game questionnaire indicated a consistent pattern with that of the donation game. In the signaling condition, the main effects of recipients' history were significant regarding the impression of the recipient ($F_{2, 102} = 90.66$, $p < .001$) and inferred goodness of the recipient's intention ($F_{2, 102} = 53.46$, $p < .001$), respectively (Figure 6a, b). A post-hoc test using Ryan's method showed that the most favorable impression and the most benign intention were attributed to the recipient with a "gave" history by participants. More importantly, the impression and goodness of intention of recipients with "did not give + abandoned" history were more favorable than those of the recipient with a "did not give + did not abandon" history.

Hypothetical behavior toward the recipient showed a similar pattern as impression and goodness of intention. The proportions of participants who chose "give" toward recipients were compared by a series of McNemar tests with the Bonferroni correction. As shown in Figure 6c, participants were more cooperative toward the recipient with the "gave" history than those with the other histories. More importantly, participants were more cooperative toward the recipient with a "did not give + abandoned" history than the recipient with the "did not give + did not abandon" history ($p < .001$ for each comparisons). These results indicated that participants interpreted the abandonment of 5 JPY after defection as a favorable behavior and a signal of benign intention to change their negative reaction toward defectors into altruistic behavior.

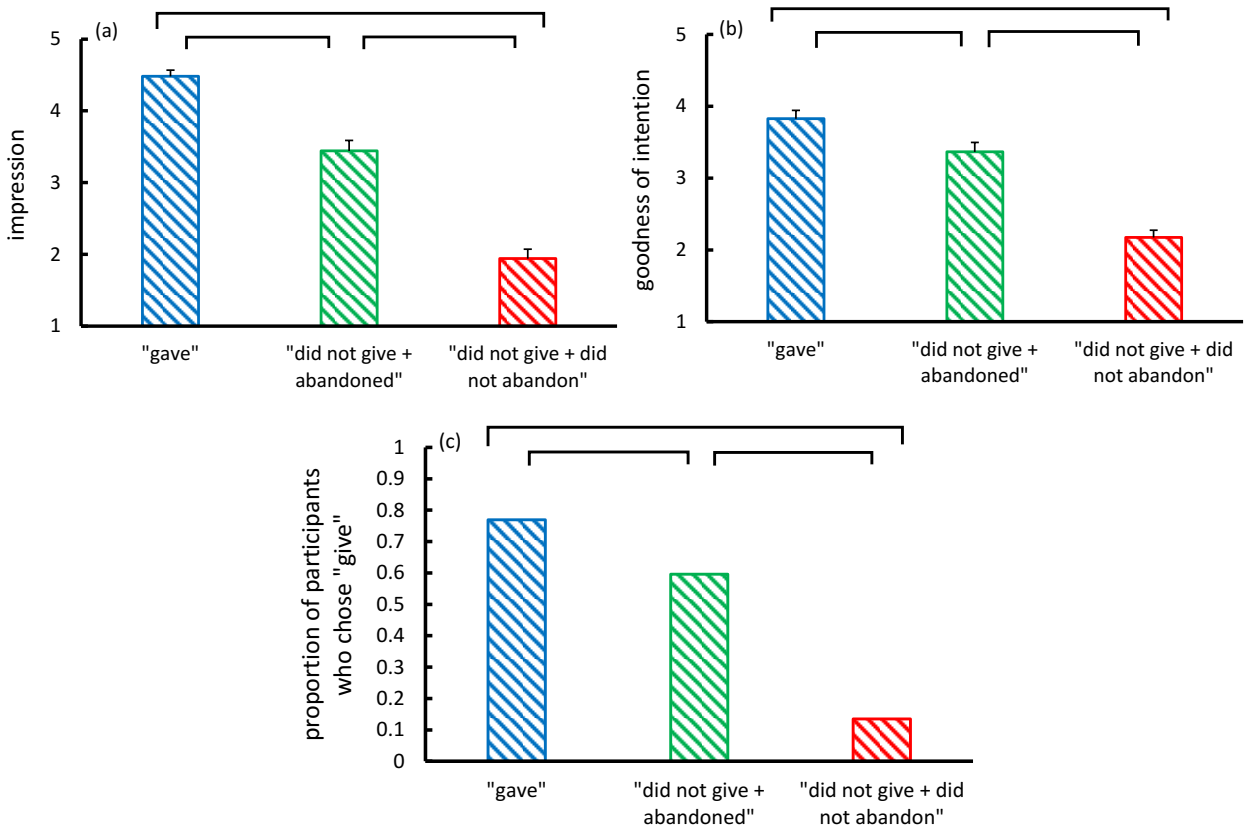

Figure 6. The results of the post-game questionnaire in the signaling session of experiment 1. The error bars indicate standard errors of the mean. (a) Mean impression score, (b) mean goodness of intention score, and (c) proportion of participants who chose "give" as a function of recipients' history.

In addition to whether to choose "give," participants' hypothetical decision about the signal option was analyzed. First, participants' willingness to choose "abandon" as a reaction to an implementation error was assessed. Fifty percent of participants reported that they would choose the "abandon" option at least once in the three situations (where the recipient was with "gave," "did not give + abandoned," and "did not give + did not abandon" histories, respectively) after committing an

implementation error. Forty-two percent of participants indicated their willingness to choose the "abandon" option in all three situations. There were 12 participants who had never chosen the "give" option for the three types of recipients. By definition, these 12 uncooperative participants would never commit implementation errors. Therefore, these participants were removed from the analyses. The recalculated proportions of participants who chose "abandon" at least once and in all three situations were both increased to 65% and 55%. Second, participants' willingness to choose "abandon" after choosing "not give" intentionally was assessed. The proportion of participants who chose "not give" toward recipients with all histories (genuine defectors) was too small to statistically compare their willingness to use the signal option with that of justified defectors, who chose "not give" only toward recipients with a "did not give + did not abandon" history. Therefore, I only report that 29% of participants who had chosen "not give" toward recipients with a "did not give + did not abandon" history, chose the "abandon" option after the "not give" choice.

In the standing condition, there were several differences between the results of the questionnaire and those of the game experiment. The main effect of the recipient history was significant, $F_{3, 153} = 94.87$, $p < .001$, and $F_{3, 153} = 41.49$, $p < .001$ for the impression of the recipient and the inferred good intention, respectively. Post-hoc tests showed that participants' impression and inferred goodness of intention were more favorable when the recipients had GG and GN histories than when recipients had NG and NN histories (Figure 7a, b). Moreover, these findings are not consistent with the game experiment in which both impression and intention scores were higher in

recipients with an NN history than in recipients with an NG history. This result suggested that participants discriminated justified defectors from unjustified defectors at impression and intention levels.

Alternatively, participants appeared not to discriminate between these two recipients at the behavioral level, even in the questionnaire (Figure 7c). A series of McNemar tests with the Bonferroni correction revealed that the proportion of participants who chose "give" toward recipients with GG and GN histories was statistically higher than those who chose "give" toward recipients with NG and NN histories *(*$p < .001$ for each comparisons). More importantly, participants' willingness to choose "give" did not differ toward the recipient with an NG history versus an NN history ($p = 1.00$). In sum of the standing condition, participants regarded the recipient with NN more favorably than the recipient with NG, while participants' hypothetical behavior toward them was not distinct from each other.

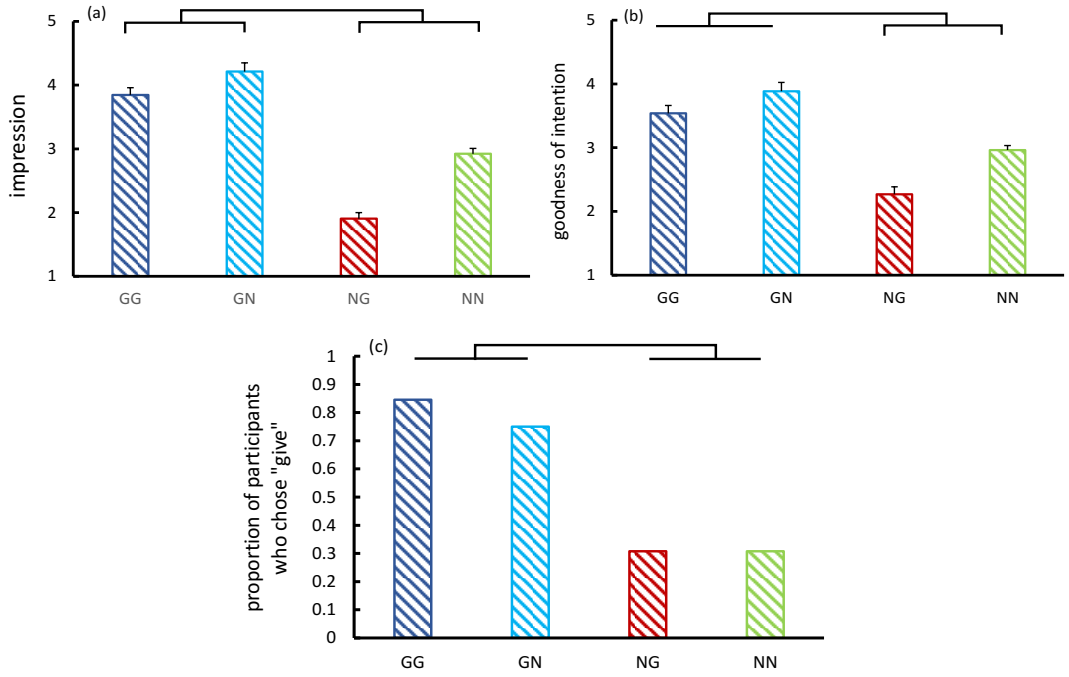

Figure 7. The results of the post-game questionnaire in the standing session of experiment 1. The error bars indicate standard errors of the mean. (a) Mean impression score, (b) mean goodness of intention score, and (c) proportion of participants who chose "give" as a function of recipients' history.

3.3. Experiment 2

To summarize the results of experiment 1, participants were likely to use the signal option when they committed an implementation error and justified defection, and they perceived other defectors' signal use as an expression of their benign intention. In spite of this supporting evidence, one limitation remains. If they were using *intSIG* precisely in an indirect reciprocity context, participants in the signaling condition of experiment 1 should have abandoned their resources after an implementation error *as frequently as* after a justified defection. In fact, participants chose "abandon" after

implementation errors *more often than* after justified defections (see Figure 2). Moreover, as shown in Figure 3, participants cooperated with cooperators ("gave") *more than* with signaling defectors ("did not give + abandoned"), which was not an ideal *intSIG*-like manner. However, it is worth noting that the donation game in experiment 1 did not present participants with obvious feedback toward their behavior. This indicates that the results in experiment 1 represented participants' natural tendency without social learning. Therefore, it is hypothesized that *including an opportunity of social learning* into the donation game would enhance their *intSIG*-like tendency shown in experiment 1.

To test this prediction, in experiment 2, participants played the donation game in the same manner as in experiment 1, except they were informed whether their partners chose "give" or "not give" in every round they were assigned to the recipient role. Moreover, this information type practically corresponded to the condition participants engaged in. In other words, participants in the signaling condition received an *intSIG*-like response, while those in the standing condition received an *ST*-like response. Thus, whether this explicit feedback would promote participants' tendency to behave in an *intSIG*- or a *ST*-like manner was tested.

Furthermore, an additional questionnaire containing the Japanese version of the Test of Self-Conscious Affect (TOSCA) was offered to participants. This questionnaire was originally made and developed by Tangney and Dearing (2002) to evaluate respondents' proneness to feel shame, guilt, and some less focal emotions. A recent study suggested that shame is an evolved psychological mechanism to limit the spread

of information detrimental to one's reputation (Sznycer et al., 2012). Additionally, Tanaka et al. (2015) showed that shame-proneness was positively correlated with people's tendency toward self-punishment, which is analogous behavior with abandoning resources demonstrated in experiment 1. Accordingly, whether shame-prone people would be likely to choose "abandon" after defection was tested.

(a) Methods

Participants. One hundred two undergraduates (48 males and 54 females) in Kobe University participated in the experiment. The mean±*SD* of participants' age was 18.79 ±0.96. The announced and actual monetary rewards for participants were the same as those of experiment 1. Three participants were dropped from the analyses, as they casted doubt on the presence of other players and requested to remove their behavioral data after debriefing.

Procedure. Here, only the procedure that differed from experiment 1 is described. In the practice session and the following signaling or standing condition, participants were informed of whether they received 10 JPY by the current donor at the end of every round they were assigned to a recipient role. If participants received 10 JPY, the displayed cumulative earnings were upgraded at the same time. The rule of the feedback was programmed as follows. First, an environment was assumed in which there is one common strategy most players use. In the experiment, out of five pre-programmed players, four players use this common strategy, and other is a genuine defector (using *ALLD*). The common strategy was determined depending on the game participants

engaged in. In the practice session, therefore, four pre-programmed players use *IS*. On the other hand, the common strategies were *intSIG* in the signaling condition, and *ST* in the standing condition. Implementation errors of pre-programmed players were not taken into account. Taking the signaling condition as an example, participants who chose "abandon" received 10 JPY in the subsequent round were assigned to the recipient role four of five times, whereby they could infer the *intSIG*-like association between their behaviors as the donor and whether they received money as the recipient. After the experimental game, participants were asked to fill out two post-game questionnaires, which were the same as experiment 1 and TOSCA. As with experiment 1, all participants were debriefed and paid 1500 JPY.

(b) Results

Cooperation rate in the practice session. In the practice session, participants' mean cooperation rate was .83 (SD = 0.22) toward the recipients with a "gave" history and .65 (SD = 0.26) toward the recipients with a "did not give" history. A 2 (recipients' history: "gave" vs. "did not give") × 2 (game type: signaling vs. standing) ANOVA revealed that the main effect of recipients' history was significant, $F_{1, 102} = 58.73$, $p < .001$. Unlike experiment 1, however, an interaction effect between the recipients' history and the game type participants engaged in was also significant, $F_{1, 97} = 5.74$, $p = .019$. Participants were more likely to choose "give" toward the recipient with a "did not give" history in the standing condition (.71, SD = 0.23) than in the signaling condition (.58, SD = 0.29). This is an unexpected result because participants were randomly

assigned to either condition. Although I could not offer any plausible explanation for this interaction, the results largely showed that participants had a tendency to behave at least in an *IS*-like manner.

Testing hypotheses about *intSIG*. Hypothesis 1a is about participants' frequency of using the signal option: Participants abandon *c* more frequently after implementation errors and justified defections than unjustified defections. The signaling rates as a function of three defection types were .70 (implementation error), .75 (justified defection), and .51 (unjustified defection), respectively (Figure 8). Fisher's exact test with the Bonferroni correction showed that participants chose "abandon" more frequently after implementation errors and justified defections than after unjustified defections ($p < .001$, for each comparison), which was a consistent pattern to that of experiment 1, supporting Hypothesis 1a. More importantly, a difference between the signaling rate after implementation errors and justified defections was not significant ($p = 0.293$). Since this difference was significant in experiment 1, the results indicate that including feedback made participants' tendency to use the signal option more similar to an *intSIG*-like manner.

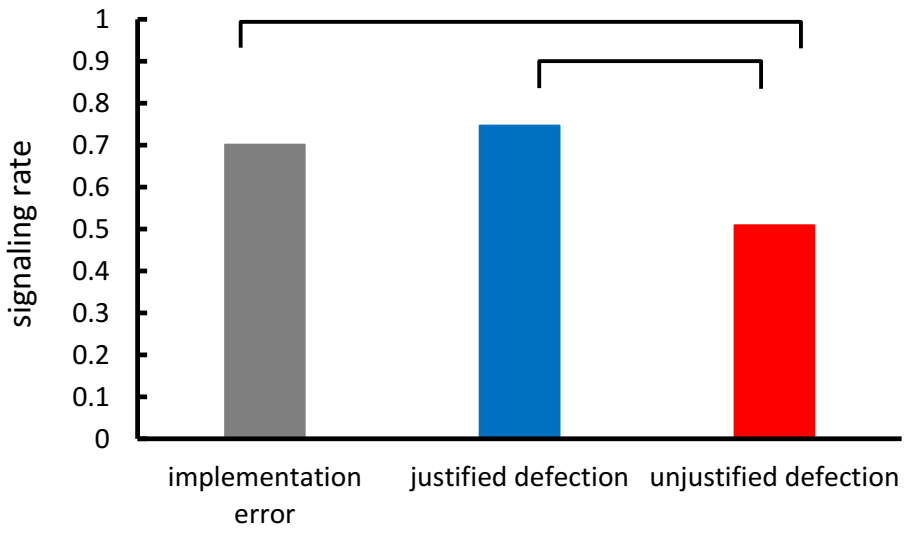

Figure 8. Signaling rate as a function of defection types in the signaling condition of experiment 2.

Hypothesis 1b is that justified defectors abandon *c* more frequently than unjustified defectors. In experiment 1, unjustified defectors were defined as participants who chose to "not give" more than 80% of the time when they were paired with recipients in good standing, while justified defectors were defined as those who were not included in unjustified defectors and chose to "not give" more than 80% of the time when they were faced with recipients in bad standing. However, in experiment 2, there were no participants corresponding to the definition of unjustified defectors. Even if 80% of the criterion was loosed to 50%, only five participants were identified as unjustified defectors. Although it was not a strict analysis because of the small sample size, the results of a two-tailed *t*-test appear to support Hypothesis 1b. The mean signaling rate of these unjustified defectors was 0.48 (SD = 0.17), which is less than that

of 18 justified defectors (0.82, $SD = 0.08$): $t_{21} = 2.14$, $p < 0.044$.

While Hypothesis 1a and 1b focused on participants' signaling behavior, Hypothesis 2 was about participants' reaction to others' signaling behavior: Participants gave *b* more frequently to other players who gave *b* or refused to give it but abandoned *c* than to players who refused both to give *b* and to abandon *c* in the previous round. As with experiment 1, the main effect of recipients' history was significant again ($F_{2,96} = 31.38$, $p < 0.001$), and a post hoc test using Ryan's method showed that participants chose "give" more frequently toward recipients with "gave" and "did not give + abandoned" histories than toward recipients with "did not give + did not abandon" history (Figure 9). More importantly, the mean cooperation rates toward recipients with "gave" and "did not give + abandoned" histories were not statistically different. This result is inconsistent with that of experiment 1, and hence not only Hypothesis 2, but also the prediction that the feedback would enhance participants' *intSIG*-like manner were supported.

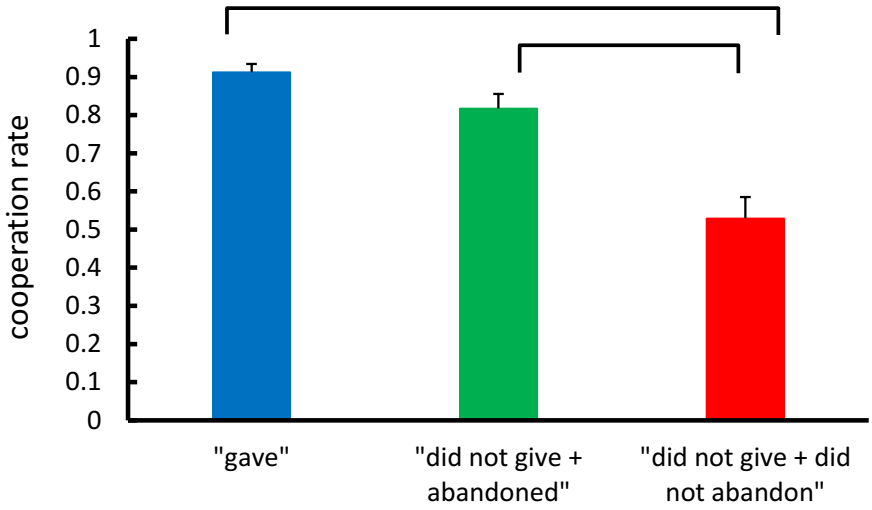

Figure 9. Cooperation rate as a function of recipients' histories in the signaling condition of experiment 1. The error bars indicate standard errors of the mean.

Testing hypotheses about *ST*. In experiment 1, participants did not discriminate justified defectors (recipients with an NN history) from unjustified defectors (recipients with an NG history). The results of experiment 2 showed a different pattern. The main effect of recipients' history was significant: $F(3,147) = 42.180$, $p < 0.001$. A post hoc test using Ryan's method showed that participants chose to "give" toward recipients with an NN history more frequently than toward recipients with an NG history (Figure 10). Although this pattern is consistent with an *ST*-like manner, the results also revealed that participants chose to "give" toward recipients with an NN history less frequently than toward recipients with GG and GN histories. This means that participants used second-information to discriminate justified defectors from unjustified defectors, but not to regard them as favorable as cooperators.

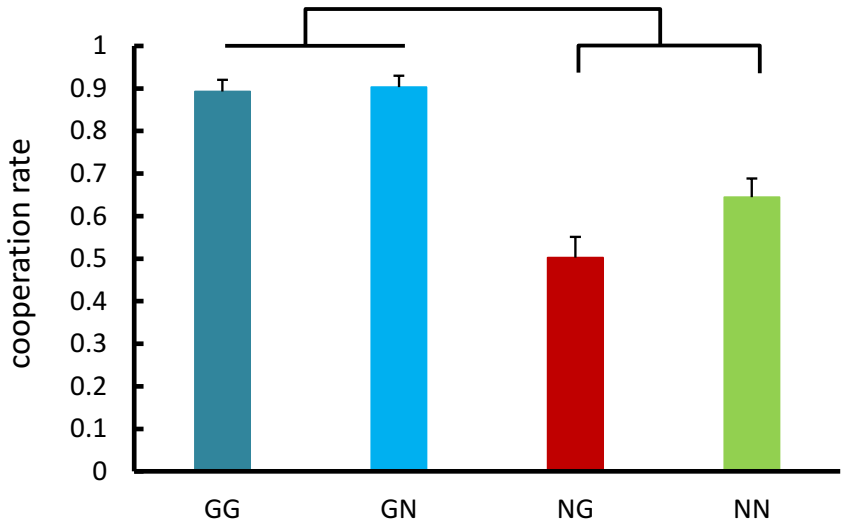

Figure 10. Cooperation rate as a function of recipients' histories in the standing condition of experiment 2. The error bars indicate standard errors of the mean.

Reaction time to recipients' history. In experiment 2, the mean RT of participants' decision of whether to choose "give" were 2.19 seconds (SD = 0.72) in the signaling condition and 2.49 seconds (SD = 0.76) in the standing condition. Unlike experiment 1, the mean RT was statistically shorter in the signaling condition than in the standing condition (t_{97} = 1.99, p = .049). This result supports the prediction that participants in the signaling condition would take less time deciding whether to choose "give" than those in the standing condition because of *ST*'s cognitive load demand. In other words, information on recipients' signal use appears less cognitively taxing than second-order information, and thus *intSIG* was easier to employ for participants than *ST*.

Participants' net payoff. Although the results of the computer simulation showed that *intSIG* is more efficient to achieve cooperative equilibrium than *ST* unless the benefit-to-cost ratio is small, it is unclear that whether the results are consistent with those of

the experiment. Therefore, the sum of the net payoffs participants earned in the two conditions was compared. Contrary to the results of the computer simulation, the mean net payoff was significantly smaller in the signaling condition (388.47 JPY, $SD = 22.32$) than in the standing condition (440.00 JPY, $SD = 42.47$): $t_{97} = 0.48$, $p = .001$, two-tailed test. However, in the experiment, participants were faced with the recipients with a "did not cooperate + did not abandon" history who were not in the simulation. This setting might provide participants in the signaling condition with more opportunities to abandon their resources, reducing their net payoff more than that of participants in the standing condition.

Post-game questionnaire. In the signaling condition, the results of the impression of the recipients and inferred goodness of recipients' intention showed identical patterns with those of experiment 1 (Figure 11a, b). The main effects of recipients' history were significant regarding the impression of the recipients ($F_{2, 96} = 150.96$, $p < .001$), and inferred goodness of recipients' intention ($F_{2, 96} = 49.40$, $p < .001$), respectively. A post-hoc test using Ryan's method indicated that the participants attributed the most favorable impression and the most benign intention to the recipient with a "gave" history, followed by "did not give + abandoned" and "did not give + did not abandon" histories in this order.

On the other hand, the hypothetical behavior toward the recipients in experiment 2 showed a different pattern from that of experiment 1 as well as impression and goodness of intention. McNemar tests with the Bonferroni correction revealed that, as shown in Figure 11c, the proportion of participants who chose "give" was greater

when the recipients had “gave” and “did not give + abandoned” histories than when the recipients had a “did not give + did not abandon” history ($p < .001$ for each comparison). More importantly, there was not a statistical difference in the proportion of participants who chose “give” between when recipients had a “gave” history versus a “did not give + abandoned” history ($p < .001$). This result is consistent with that of the signaling condition of experiment 2 in which participants regarded signaling defectors as favorably as cooperators.

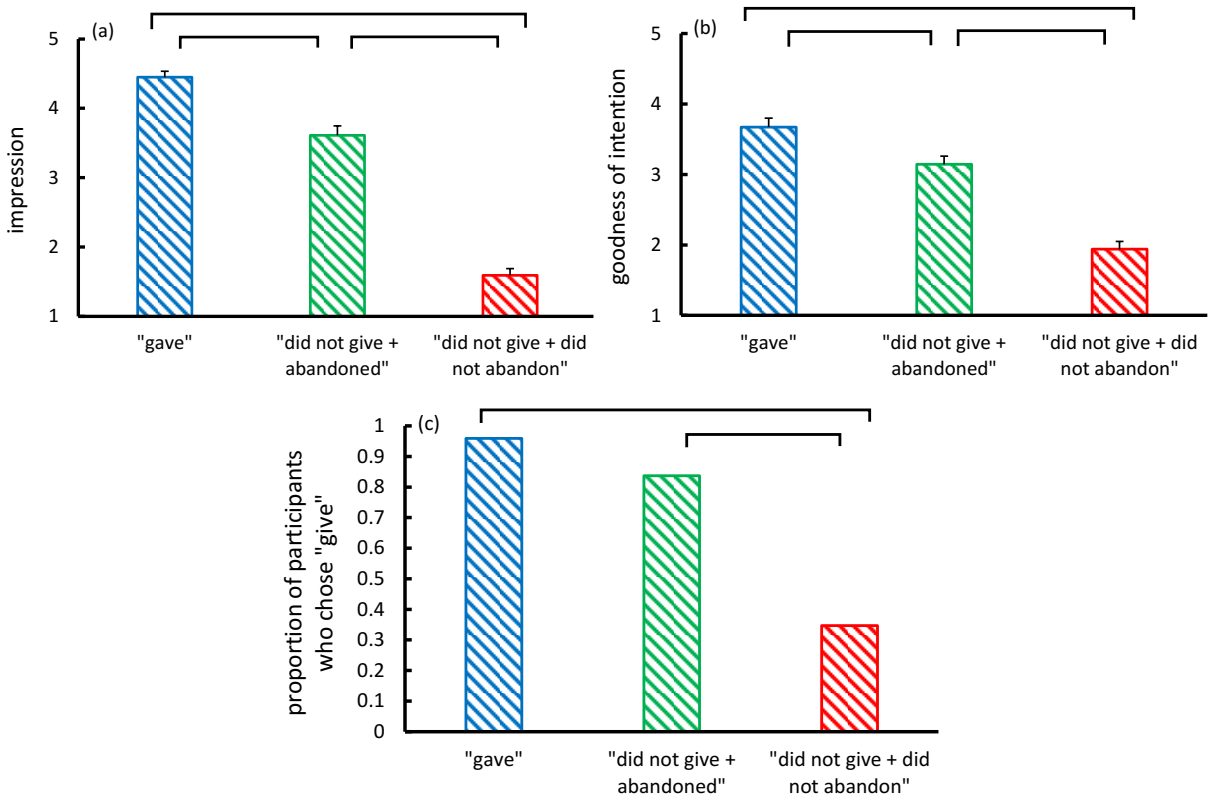

Figure 11. The results of the post-game questionnaire in the signaling session of experiment 2. The error bars indicate standard errors of the mean. (a) Mean impression score, (b) mean goodness of intention score, and (c) proportion of participants who chose “give” as a function of recipients’ history.

Regarding participants' willingness to use the signal option after committing an implementation error, 82% of participants noted that they would choose "abandon" at least once in three situations of recipients' histories ("gave," "did not give + abandoned," and "did not give + did not abandon") after committing an implementation error. The proportion of participants who always chose "abandon" was 76%. In experiment 1, these proportions were computed again, removing the data of genuine defectors who never committed an implementation error. However, in experiment 2, there was only one participant who had never chosen "give," so this re-calculation was not employed. Similarly, it was not tested whether justified defectors, who chose "not give" only toward the recipient with a "did not give + did not abandon" history, were more likely to choose "abandon" than genuine defectors, who chose "not give" toward all recipients. The reason is the same as experiment 1, as the proportion of genuine defectors were small. I only reported here that the proportion of justified defectors who chose "abandon" was 90%.

In the standing condition, the results of all items (the impression of recipients, inferred goodness of recipients' intention, and hypothetical behavior toward recipients) showed identical patterns to experiment 1 (Figure 12). As for the impression and inferred goodness of intention of recipients, the main effect of the recipient history was significant, $F_{3, 147} = 111.46$, $p < .001$, and $F_{3, 147} = 58.27$, $p < .001$, respectively. Post-hoc tests using Ryan's method indicated that participants' impression and the inferred goodness of intention of recipients with GG and GN histories were more favorable than

recipients with NG and NN histories. More importantly, both impression and intention scores were rated higher when recipients had an NN history than when recipients had an NG history. Nevertheless, participants' hypothetical behavior toward the recipients with these histories was equally favorable. A series of McNemar tests with the Bonferroni correction showed that the proportions of participants who chose "give" toward recipients with NG and NN histories were not statistically different ($p = 0.052$). Consistent with experiment 1, although participants might attribute different levels of goodness to the recipients with NG and NN histories as different, they appear to react to them identically at the behavioral level.

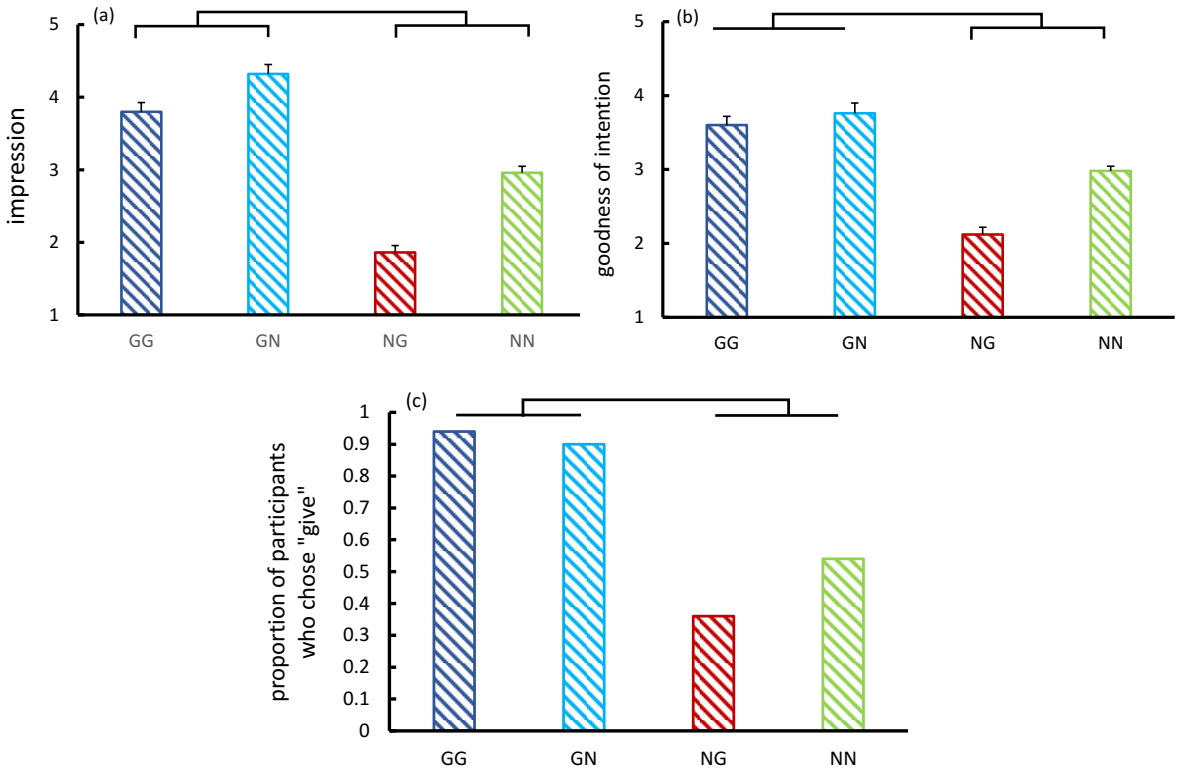

Figure 12. The results of the post-game questionnaire in the standing session of experiment 2. The error bars indicate standard errors of the mean. (a) Mean impression score, (b) mean goodness of intention score, and (c) proportion of participants who chose "give" as a function of recipients' history.

In experiment 2, participants were further asked to fill out the TOSCA, which measures respondents' proneness to feel shame and guilt. Although it was predicted that participants' shame-proneness would positively correlate with their tendency to use the signal option, no participants' behavioral variable in the donation game correlated to the score of trait shame and trait guilt. This could be because, in the donation game of present study, participants could choose a behavior related to the signal from only

binary options (either “abandon” or “not abandon”). In other words, the amount of the signal cost in the experiment was fixed. However, in reality, the signal cost to communicate benign intention is continuous. If participants choose the amount of the signal cost, this amount may correlate with shame-proneness.

Chapter 4

Discussion

In the two experiments, the availability of the signal option evidently influenced participants' behavior. They voluntarily abandoned their resources after committing a defection. This tendency was found even when they were not aware of whether other players would favorably react to it (experiment 1), and when they had some clues about other players' favorable reaction to the signal, the tendency was facilitated further (experiment 2). It cannot be overemphasized that participants used the signal in spite of *its costliness*. In a dyadic relationship, it has been proven that people displayed costly behavior when they unintentionally defected on their partner (Watanabe & Ohtsubo, 2012). The results of the present experiments that participants abandoned their resources after an unintentional defection is not only consistent with this study, but also first demonstrated this costly behavior in an indirect reciprocity context, which is beyond dyadic interaction. More surprisingly, the present experiments also showed that participants emitted a costly signal when they *intentionally* defected against unjustified defector. Of course, they might do so as a righteous reaction to injustice, but is there a necessity for such a person to voluntarily abandon her/his resources? There should be if other people do not recognize that she/he behaved righteously. This means that participants did not use the signal option simply because they intended to express their *remorse* for defection. Rather, they informed others of their *benign intention*. This suggestion is also supported by another result that

participants, as signal receivers, attributed more benign intention to signaling defectors than non-signaling defectors and cooperated more frequently with the former. Since the evolutionary game analysis and the computer simulation undergird aspects of mathematical robustness, it can be concluded that *intSIG* has both theoretical and empirical validities as a strategy for indirect reciprocity. However, this study possesses the limitation that participants played games with pre-programmed pseudo players. Therefore, whether *intSIG* spontaneously emerges in real interaction remains to be examined.

4.1. The signal option vs. second-order information

As for the standing strategy (*ST*), participants did not used second-order information in experiment 1, while they used it to some extent in experiment 2. This indicates that participants in experiment 2 learned to behave in an *ST*-like manner from the repetitive *ST*-like feedback toward their behavior. In other words, participants use second-order information as long as other players also use it. This indicates that if people naturally behave, as shown in experiment 1, *ST* is less likely to emerge in real interactions than *intSIG* because people never receive *ST*-like feedback by others who also do not use second-order information prior to learning.

This result is not consistent with recent findings that participants tended to request second-order information involved with their partners' previous defection (Swakman et al., 2016). However, Swakman et al.'s study also showed that the frequency of participants who requested the second-order information and actually behaved in an

ST-like manner is smaller than the frequency of those who either did not request the second-order information or requested it but failed to adequately use it. One plausible speculation is that people unconsciously know that second-order information is useful to assess a defector's intention but it is tough for most of them to interpret such information adequately. Although *ST* regards players who defected on a recipient in bad standing as a justified defector, by definition, a set of justified defectors includes unconditional defectors (who refuse to cooperate regardless of recipients' standing). Moreover, second-order information is less reliable than first-order information because it is less recent than first-order information so that it can be distorted through transmission (Panchanathan, 2011; Swakman et al., 2016). Therefore, second-order information may be indicative of defectors' intention to some extent, but not enough to determine it precisely.

In addition to empirical findings, the results of the simulation study showed that *ST* is less efficient than *intSIG* under cooperative equilibrium unless the benefit-to-cost ratio is small. This is due to the quickness of *intSIG*'s reputation recovery when players defect. As strategies without a signal option, *ST* should turn a blind eye to defection against others' unintentional failure to cooperate. Therefore, the cooperation rate under cooperative equilibrium would be truncated by this "justified" (by a *ST*-like manner) defection, which never occurred in the *intSIG* group. It is noteworthy that these implications apply not only to the standing strategy, but also to every other strategy that uses second-order information (Kandori, 1992; Ohtsuki & Iwasa, 2004; 2006; Takahashi & Mashima, 2003).

4.2. The demanded amount of the signal cost in an empirical context

Although the present study confirmed that people behaved in an *intSIG*-like manner, there is another empirical question: what *amount* of the resource are they actually willing to abandon for their reputation? In the evolutionary game analysis, it is assumed that the cost of the signal must be at least equal to the cost of cooperation, otherwise free-riders can fake the signal at a lower cost. Based on this assumption, in experiments 1 and 2, the cost of the signal option was fixed at five JPY, which was equivalent to the cooperation cost. Accordingly, if participants can choose not only whether to use the signal option, but also what amount of the cost they pay for the signal option, the result could deviate from the theoretical expectations.

At this stage, I do not have a clear-cut prediction regarding this point. According to the present study, however, people learn to behave precisely by others' reaction. This suggests that the amount of the cost they pay for the signal option would depend on a criterion other people use to infer defectors' intention. Therefore, if a group contains people who are tolerant to low-cost signals, the net amount of the signal cost at a group level would be less than the amount of the cooperation cost. Contrary, if there are strict signal receivers in the group, a maladaptive consequence of signal inflation would result. One possible example of this phenomenon is suicide, which may occur as an unfortunate byproduct of costly signaling (Syme, Garfield, & Hagen, 2015). If the signal inflation can be demonstrated in the donation game by manipulating group-level signal criteria, the further research on *intSIG* will have applicable implications for some practical problems.

4.3. Intention signaling as an additional behavioral option

In the present study, what was proposed to maintain cooperative equilibrium is introducing intention signaling in a repertoire of players' behavior. As traditional indirect reciprocity research has assumed it to be a binary option (either cooperation or defection), *intSIG* might appear to be far from such an ordinal approach. To broaden the horizon, however, there are similar theoretical approaches that also expand the behavioral option to understand the evolution of cooperation.

Typical options previous studies have proposed are withdrawing from uncooperative players or seeking other favorable partners (Barclay & Raihani, 2016; Barclay & Willer, 2007; Fu, Hauert, Nowak, & Wang, 2008; Hauert, De Monte, Hofbauer, & Sigmund, 2002; Hauert, Traulsen, Brandt, Nowak, & Sigmund, 2007; Noë & Hammerstein, 1994; Pacheco, Traulsen, & Nowak, 2006; Wardil & Hauert, 2014). Some recent literature has attempted to introduce such options into the indirect reciprocity context. In Ghang and Nowak's model (2015), each player can decide whether to interact with her/his current partner prior to deciding whether to cooperate. Withdrawing from interactions with uncooperative players does not sully cooperators' reputation. Regarding partner choice, Roberts (2015) showed that if each donor was allowed to keep searching for a partner until she/he met one whose image score satisfied her/his criterion, levels of cooperation were higher and more stable than if players were randomly paired. Although these options are different from *intSIG*, all seem to converge on a common theme that *only deciding whether to cooperate is not sufficient*. An

additional behavioral option to cooperation/defection enables players to distinguish their apparently uncooperative behaviors (e.g., not giving a resource to bad players) from genuinely uncooperative behaviors. Only if there is no incentive for genuine defectors to use it, any behavior can function to distinguish justified and unjustified defectors. The present study set such disincentive as the cost of using the option.

4.4. How can signals emerge?

The evolutionary game analysis and the simulation study theoretically support the viability of *intSIG* as strategy for indirect reciprocity. However, both studies were conducted under an assumption that every player has an adaptive set of behavioral (signaling propensity) and reputation-assignment strategies (propensity to adequately respond to the signal). Although a group composed of *intSIG* players is stable against unconditional defector/cooperator and achieves a high level of cooperation, it is still unclear *how such propensities can emerge in the first place*.

One possibility is that signaling and signal reading first co-evolved in a direct reciprocity context. Mutual cooperation sustained by tit-for-tat players in the iterated prisoner's dilemma is vulnerable to even one implementation error, inducing an endless retaliation for partner's defection (Nowak & Sigmund, 1992). Immediate communication of a careless defector's benign intention could allow tit-for-tat players to prevent such unfortunate retaliation. This prevention is beneficial for both the signaler and signal reader. Moreover, when players can voluntarily withdraw from the uncooperative interaction, a costly signal after an implementation error could also

prevent the premature dissolution of potentially beneficial, long-term relationships (Ohtsubo & Watanabe, 2009). Therefore, if a once costly signal emerges in the direct reciprocity context, it can be applied to the indirect reciprocity context. Another possibility is that the ability of signal communication is an evolutionary tributary originating from partner choice. Unlike indirect reciprocity, where there is a cost associated with helping "good" players, choosy players in a partner choice context do not have to incur the cost of choosiness (Barclay, 2013; Raihani & Bshary, 2016; Sylwester & Roberts, 2013). Accordingly, the cost-perceiving ability to seek other players' cooperativeness might have first evolved, accompanied by the emergence of active cost-incurring behavior. Although the emergence of signal communication seems to be a difficult question beyond the theme of this paper, once evolved in some domain, it might have been expanded to the indirect reciprocity context. Rather, if human communication emerged from social skills, such as cooperation with others (Dunbar, 1998; Humphrey, 1976; Tomasello, 2010), the theory of *intSIG* may provide an evolutionary explanation of human traits with a new path.

4.5. Conclusion: Human beings are not only a cooperative, but also a communicative species

As noted, a costly signal is just one of many possible behavioral options to discriminate justified defectors from unjustified defectors. However, it has a profound, specific implication for human sociality.

There are several theories explaining the evolution of large-scale human

cooperation other than indirect reciprocity. Of such alternative explanations, the most influential models were *competitive altruism* (Roberts, 1998, Barclay, 2006) and *strong reciprocity* (Gintis, 2000). According to the theory of competitive altruism, a cost associated with cooperation signals actors' cooperativeness, then they can receive good reputation eliciting others' help. On the other hand, the theory of strong reciprocity focuses on the importance of punishment against defectors to reduce their fitness. Although the evolution of punishment remains as a question because of its costliness for punishers, recent literature argued that this costliness can function as a signal of trustworthiness (Jordan, Hoffman, Bloom, & Rand, 2016). In sum, it has already been suggested that a costly signal plays a crucial role in human cooperation.

Nevertheless, the implications of the present study are largely distinct from these theories because they conceptualize *altruistic behaviors (cooperation with good one*s*, and punishment against bad one*s*) themselves as signals*. According to this conceptualization, to make bonds, all we humans have to do is to behave altruistically. However, it has been documented that many apparently wasteful behaviors, which cannot be equated with altruistic behaviors, also serve as commitment signals and facilitate dyadic cooperation by cementing interpersonal bonds (Frank, 1988; Nesse, 2001; Yamaguchi, Smith, & Ohtsubo, 2015). The *intSIG* strategy likewise incorporates a signaling option independent of cooperation and allows players to maintain their good standing even when they withhold help. This idea is resonant with the notion of communicative cooperation coined by Noë (2006). Although it was proposed to underscore the importance of communication in animal cooperation, communications

via signals should be no less important for human beings, as we are not only a highly *cooperative* species but also an extremely *communicative* one. The present study shed light on the latter aspects of human beings to solve an enduring enigma, which is evolution of large-scale human cooperation.

[Chapter 1, a part of Chapter 2 (except 2.3), Chapter 3 and a part of Chapter 4 (except 4.1 and 4.2) is based on Tanaka, H., Ohtsuki, H., & Ohtsubo, Y. (2016). The price of being seen to be just: an intention signalling strategy for indirect reciprocity. *Proceedings of the Royal Society B*, *283*(1835), 20160694.]

References

Alexander, R. D. (1987). *The biology of moral systems*. New York, NY: Aldine de Gruyter.

Axelrod, R. (1984). *The evolution of cooperation*. New York, NY: Basic Books.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(4489), 1390-1396. doi:10.1126/science.7466396

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior, 34*(3), 164-175. doi:10.1016/j.evolhumbehav.2013.02.002

Barclay, P., & Raihani, N. (2016). Partner choice versus punishment in human Prisoner's Dilemmas. *Evolution and Human Behavior, 37*(4), 263-271. doi:10.1016/j.evolhumbehav.2015.12.004

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B, 274*(1610), 749-753. doi:10.1098/rspb.2006.0209

Batra, S. W. (1968). Behavior of some social and solitary halictine bees within their nests: A comparative study. *Journal of the Kansas Entomological Society, 41*(1), 120-133.

Bolton, G. E., Katok, E., & Ockenfels, A. (2004). How effective are electronic reputation mechanisms? An experimental investigation. *Management science,*

50(11), 1587-1602. doi:10.1287/mnsc.1030.0199

Bolton, G. E., Katok, E., & Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Public Economics, 89*(8), 1457-1468. doi:10.1016/j.jpubeco.2004.03.008

Bowles, S., & Gintis, H. (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton, NJ: Princeton University Press.

Boyd, R. (1989). Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *Journal of Theoretical Biology, 136*(1), 47-56. doi:10.1016/S0022-5193(89)80188-2

Brown, J. L. (1978). Avian communal breeding systems. *Annual Review of Ecology and Systematics, 9*, 123-155. doi:10.1146/annurev.es.09.110178.001011

Crespi, B. J., & Yanega, D. (1995). The definition of eusociality. *Behavioral Ecology, 6*(1), 109-115. doi:10.1093/beheco/6.1.109

Dufwenberg, M., Gneezy, U., Güth, W., & Van Damme, E. (2001). Direct vs indirect reciprocity: An experiment. *Homo Oeconomics, 18*, 19-30.

Dunbar, R. (1998). *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard University Press.

Engelmann, D., & Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior, 67*(2), 399-407. doi:10.1016/j.geb.2008.12.006

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: a field experiment. *Evolution and Human Behavior, 32*(3),

172-178. doi:10.1016/j.evolhumbehav.2010.10.006

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*, 785-791. doi:doi:10.1038/nature02043

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture*. New York, NY: McGraw-Hill.

Frank, R. H. (1988). *Passions within reason: the strategic role of the emotions*. New York, NY: Norton.

Fu, F., Hauert, C., Nowak, M. A., & Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Physical Review E, 78*(2), 026117. doi:10.1103/PhysRevE.78.026117

Ghang, W., & Nowak, M. A. (2015). Indirect reciprocity with optional interactions. *Journal of Theoretical Biology, 365*, 1-11. doi:10.1016/j.jtbi.2014.09.036

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology, 206*(2), 169-179. doi:10.1006/jtbi.2000.2111

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology, 144*(4), 517-546. doi:10.1016/S0022-5193(05)80088-8

Hamilton, W. D. (1964). The genetical evolution of social behaviour, I & II. *Journal of Theoretical Biology, 7*(1), 1-52. doi:10.1016/0022-5193(64)90038-4, 10.1016/0022-5193(64)90039-6

Hatchwell, B. J., Russell, A. F., MacColl, A. D., Ross, D. J., Fowlie, M. K., & McGowan, A. (2004). Helpers increase long-term but not short-term productivity in cooperatively breeding long-tailed tits. *Behavioral Ecology,*

15(1), 1-10. doi:10.1093/beheco/arg091

Hatchwell, B. J., & Sharp, S. P. (2006). Kin Selection, Constraints, and the Evolution of Cooperative Breeding in Long‐Tailed Tits. *Advances in the Study of Behavior, 36*, 355-395. doi:10.1016/S0065-3454(06)36008-1

Hauert, C., De Monte, S., Hofbauer, J., & Sigmund, K. (2002). Volunteering as red queen mechanism for cooperation in public goods games. *Science, 296*(5570), 1129-1132. doi:10.1126/science.1070582

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: the emergence of costly punishment. *Science, 316*(5833), 1905-1907. doi:10.1126/science.1141588

Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing points in ethology* (pp. 303-317). Cambridge, UK: Cambridge University Press.

Jacquet, J., Hauert, C., Traulsen, A., & Milinski, M. (2012). Could shame and honor save cooperation? *Communicative & integrative biology, 5*(2), 209-213. doi:10.4161/cib.19016

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*(7591), 473-476. doi:10.1038/nature16981

Kandori, M. (1992). Social norms and community enforcement. *The Review of Economic Studies, 59*(1), 63-80. doi:10.2307/2297925

Kato-Shimizu, M., Onishi, K., Kanazawa, T., & Hinobayashi, T. (2013). Preschool

children's behavioral tendency toward social indirect reciprocity. *PloS one, 8*(8), e70915. doi:10.1371/journal.pone.0070915

Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society B, 268*(1468), 745-753. doi:10.1098/rspb.2000.1573

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge, UK: Cambridge university press.

Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature, 246*(5427), 15. doi:10.1038/246015a0

Michener, C. D. (1969). Comparative social behavior of bees. *Annual review of entomology, 14*(1), 299-342. doi:10.1146/annurev.en.14.010169.001503

Milinski, M., Semmann, D., Bakker, T. C., & Krambeck, H. J. (2001). Cooperation through indirect reciprocity: image scoring or standing strategy? *Proceedings of the Royal Society B, 268*(1484), 2495-2501. doi:10.1098/rspb.2001.1809

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002b). Reputation helps solve the 'tragedy of the commons'. *Nature, 415*(6870), 424-426. doi:10.1038/415424a

Milinski, M., Semmann, D., & Krambeck, H. (2002a). Donors to charity gain in both indirect reciprocity and political reputation. *Proceedings of the Royal Society B, 269*(1494), 881-883. doi:10.1098/rspb.2002.1964

Nesse, R. (2001). *Evolution and the capacity for commitment*. New York, NY: Russell Sage Foundation.

Noë, R. (2006). Cooperation experiments: Coordination through communication versus

acting apart together. *Animal Behaviour, 71*(1), 1-18. doi:10.1016/j.anbehav.2005.03.037

Noë, R., & Hammerstein, P. (1994). Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral ecology and sociobiology, 35*(1), 1-11. doi:10.1007/BF00167053

Nowak, M. A. (2012). Evolving cooperation. *Journal of Theoretical Biology, 299*, 1-8. doi:10.1016/j.jtbi.2012.01.014

Nowak, M. A., & Highfield, R. (2011). *Supercooperators: Altruism, evolution, and why we need each other to succeed*. New York, NY: Free Press.

Nowak, M. A., & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature, 355*(6357), 250-253. doi:10.1038/355250a0

Nowak, M. A., & Sigmund, K. (1998a). The dynamics of indirect reciprocity. *Journal of Theoretical Biology, 194*(4), 561-574. doi:10.1006/jtbi.1998.0775

Nowak, M. A., & Sigmund, K. (1998b). Evolution of indirect reciprocity by image scoring. *Nature, 393*(6685), 573-577. doi:10.1038/31225

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291-1298. doi:10.1038/nature04131

Ohtsubo, Y., & Rapoport, A. (2006). Depth of reasoning in strategic form games. *Journal of Socio-Economics, 35*(1), 31-47. doi:10.1016/j.socec.2005.12.003

Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior, 30*(2), 114-123. doi:10.1016/j.evolhumbehav.2008.09.004

Ohtsuki, H., & Iwasa, Y. (2004). How should we define goodness?--reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology, 231*(1), 107-120. doi:10.1016/j.jtbi.2004.06.005

Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology, 239*(4), 435-444. doi:10.1016/j.jtbi.2005.08.008

Pacheco, J. M., Traulsen, A., & Nowak, M. A. (2006). Active linking in evolutionary games. *Journal of Theoretical Biology, 243*(3), 437-443. doi:10.1016/j.jtbi.2006.06.027

Panchanathan, K. (2011). Two wrongs don't make a right: The initial viability of different assessment rules in the evolution of indirect reciprocity. *Journal of Theoretical Biology, 277*(1), 48-54. doi:10.1016/j.jtbi.2011.02.009

Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology, 224*(1), 115-126. doi:10.1016/s0022-5193(03)00154-1

Raihani, N. J., & Bshary, R. (2015). Third‐party punishers are rewarded, but third‐party helpers even more so. *Evolution, 69*(4), 993-1003. doi:10.1111/evo.12637

Roberts, G. (1998). Competitive altruism: From reciprocity to the handicap principle. *Proceedings of the Royal Society B, 265*(1394), 427-431. doi:10.1098/rspb.1998.0312

Roberts, G. (2015). Partner choice drives the evolution of cooperation via indirect reciprocity. *PloS one, 10*(6), e0129442. doi:10.1371/journal.pone.0129442

Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature, 444*(7120), 718-723. doi:10.1038/nature05229

Seinen, I., & Schram, A. (2006). Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European Economic Review, 50*(3), 581-602. doi:10.1016/j.euroecorev.2004.10.005

Seyfarth, R. M., & Cheney, D. L. (1984). Grooming, alliances and reciprocal altruism in vervet monkeys. *Nature, 308*(5959), 541-543. doi:10.1038/308541a0

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the national academy of sciences, 104*(44), 17435-17440. doi:10.1073/pnas.0704598104

Sugden, R. (1986). *The economics of rights, co-operation and welfare*. Oxford, UK: Blackwell.

Swakman, V., Molleman, L., Ule, A., & Egas, M. (2016). Reputation-based cooperation: empirical evidence for behavioral strategies. *Evolution and Human Behavior, 37*(3), 230-235. doi:10.1016/j.evolhumbehav.2015.12.001

Sylwester, K., & Roberts, G. (2013). Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evolution and Human Behavior, 34*(3), 201-206. doi:10.1016/j.evolhumbehav.2012.11.009

Syme, K. L., Garfield, Z. H., & Hagen, E. H. (2016). Testing the bargaining vs. inclusive fitness models of suicidal behavior against the ethnographic record. *Evolution and Human Behavior, 37*(3), 179-192.

doi:10.1016/j.evolhumbehav.2015.10.005

Sznycer, D., Takemura, K., Delton, A. W., Sato, K., Robertson, T., Cosmides, L., & Tooby, J. (2012). Cross-cultural differences and similarities in proneness to shame: An adaptationist and ecological approach. *Evolutionary Psychology, 10*(2), 352-370.

Takahashi, N., & Mashima, R. (2003). The emergence of indirect reciprocity: Is the standing strategy the answer. *Center for the study of cultural and ecological foundations of the mind, Hokkaido University, Japan, Working paper series, 29*.

Tanaka, H., Yagi, A., Komiya, A., Mifune, N., & Ohtsubo, Y. (2015). Shame-prone people are more likely to punish themselves: A test of the reputation-maintenance explanation for self-punishment. *Evolutionary Behavioral Sciences, 9*(1), 1-7. doi:10.1037/ebs0000016

Tangney, J. P., & Dearing, R. L. (2003). *Shame and guilt*. New York, NY: Guilford Press.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie, 20*(4), 410-433. doi:10.1111/j.1439-0310.1963.tb01161.x

Tomasello, M. (2010). *Origins of human communication*: MIT press.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35-57. doi:10.1086/406755

Ule, A., Schram, A., Riedl, A., & Cason, T. N. (2009). Indirect punishment and generosity toward strangers. *Science, 326*(5960), 1701-1704. doi:10.1126/science.1178883

Wardil, L., & Hauert, C. (2014). Origin and structure of dynamic cooperative networks. *Scientific reports, 4*(5725). doi:10.1038/srep05725

Watanabe, E., & Ohtsubo, Y. (2012). Costly apology and self-punishment after an unintentional transgression. *Journal of Evolutionary Psychology, 10*(3), 87-105. doi:10.1556/jep.10.2012.3.1

Wedekind, C., & Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Current Biology, 12*(12), 1012-1015. doi:10.1016/S0960-9822(02)00890-4

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science, 288*(5467), 850-852. doi:10.1126/science.288.5467.850

Wilkinson, G. S. (1984). Reciprocal food sharing in the vampire bat. *Nature, 308*(5955), 181-184. doi:10.1038/308181a0

Wilkinson, G. S. (1988). Reciprocal altruism in bats and other mammals. *Ethology and Sociobiology, 9*(2-4), 85-100. doi:10.1016/0162-3095(88)90015-5

Yamaguchi, M., Smith, A., & Ohtsubo, Y. (2015). Commitment signals in friendship and romantic relationships. *Evolution and Human Behavior, 36*(6), 467-474. doi:10.1016/j.evolhumbehav.2015.05.002

Yoeli, E., Hoffman, M., Rand, D. G., & Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the national academy of sciences, 110*(Supplement 2), 10424-10429. doi:10.1073/pnas.1301210110

Zahavi, A., & Zahavi, A. (1997). *The handicap principle: A missing part of Darwin's*

puzzle. New York, NY: Oxford University Press.

Acknowledgements

I owe an enourmous debt of gratitude to Associate Professor Yohsuke Ohtsubo for his generosity and patient support throughout all my study. I also very grateful for technical assistance of theoretical part of this study given by Assistance Professor Hisashi Ohtsuki. Associate Professor Keiko Ishii, Professor Shinichi Kita, Professor Ichiro Nagasaka, and Associate Professor Yasuki Noguchi provided me with very valuable and constructive advice. Assistant and encouragement provided by my laboratory members, Adam Smith, Toshiyuki Himichi, Ayano Yagi, Mana Yamaguchi, Ye-Yun Yu, Naoki Konishi, Keisuke Matsugasaki, and Chiaki Yamaguchi were very much appreciated.

Publications

【International jounals (peer review)】

Tanaka, H., Ohtsuki, H., & Ohtsubo, Y. (2016). The price of being seen to be just: an intention signalling strategy for indirect reciprocity. *Proceedings of the Royal Society B*, *283*(1835), 20160694.

Tanaka, H., Yagi, A., Komiya, A., Mifune, N., & Ohtsubo, Y. (2015). Shame-prone people are more likely to punish themselves: A test of reputation-maintenance explanation for self-punishment. *Evolutionary Behavioral Sciences*, *9*(1), 1-7, 2015.

Ohtsubo, Y., Matsunaga, M., Komiya A., Tanaka, H., Mifune, N., & Yagi, A. (2014). Oxytocin receptor gene (OXTR) polymorphism and self-punishment after an unintentional transgression. *Personality and Individual Differences*, ***69***, 182-186.

【Domestic jounals (peer review)】

Tanaka, H. (2013). The effect of intention in an ultimatum game using a game method. Japanese Journal of Social Psychology, *29*(1)**,** 21 – 24.

【International conferences】

Tanaka, H. (2016, July). Punishers' paying costs to maintain their reputation preserves large-scale human cooperation from collapse. 31st International Congress of Psychology, Yokohama, Japan.

Tanaka, H., Ohtsuki, H., & Ohtsubo, Y. (2016, June). Costly signaling leads to a more efficient cooperative equilibrium than the standing strategy. Human Behavior & Evolution Society 26th Annual Conference, Vancouver, Canada.

Tanaka, H., & Ohtsubo, Y. (2015, May). Punishers pay costs to maintain their reputation in the context of indirect reciprocity. Human Behavior & Evolution Society 25th Annual Conference, Columbia, Missouri.

Tanaka, H., Komiya, A., Mifune, N., Yagi, Ayano., & Ohtsubo, Y. (2013, July). Shame-prone People Are More Likely to Punish Themselves. Human Behavior & Evolution Society 25th Annual Conference, Miami, Florida.

【Domestic Conferences】

大坪庸介・松永昌宏・田中大貴・小林章雄・柴田英治・堀礼子・梅村朋弘・大平英樹. (2016 年 9 月).「ごめんね」だけでは誠意は伝わらない: 謝罪-赦しに関する fMRI 研究. 日本社会心理学会第 57 回大会, 関西学院大学.

Tanaka, H., Ohtsuki, H., & Ohtsubo, Y. (2015 年 12 月). The evolutionary stability of the intention signaling strategy for indirect reciprocity. 日本人間行動進化学会第８回大会, 総合研究大学院大学.

田中大貴・Eric Pedersen・Michael McCullough・大坪庸介. (2015 年 11 月). 赤の他人のもめごとにわざわざ関わる人はいるのか？ 第三者罰・介入の外

的妥当性に関する研究. 日本社会心理学会第 56 回大会, 東京女子大学.

Tanaka, H., & Ohtsubo, Y. (2015 年 10 月). Signaling benign intention in indirect reciprocity: How can punishers manage their own impression? 日本グループ・ダイナミックス学会第 62 回大会, 奈良大学.

田中大貴・大坪庸介. (2014 年 11 月). 間接互恵状況におけるシグナルの授受の共進化の実験的検討. 日本人間行動進化学会第 7 回大会, 神戸大学.

田中大貴・大坪庸介. (2014 年 7 月). 意図のシグナルが懲罰的非協力者の評判に及ぼす効果. 日本社会心理学第 55 回大会, 北海道大学.

田中大貴・大坪庸介. (2013 年 12 月). 間接互恵状況における協力意図のシグナル. 日本人間行動進化学会第 6 回大会, 広島修道大学.

大坪庸介・松永昌宏・田中大貴・八木彩乃・三船恒裕・小宮あすか. (2013 年 11 月). 自己罰傾向とオキシトシン受容体遺伝子多型. 日本社会心理学会第 54 回大会, 沖縄国際大学.

田中大貴・小宮あすか・三船恒裕・八木彩乃・大坪庸介. (2013 年 11 月). 自己罰行動の至近要因としての恥感情. 日本社会心理学会第 54 回大会, 沖縄国際大学.

田中大貴・小宮あすか・三船恒裕・八木彩乃・大坪庸介. (2012 年 12 月). 個人特性としての恥感情(trait shame)と自己罰行動との関連. 日本人間行動進化学会第 5 回大会, 東京大学.

田中大貴. (2012 年 11 月). ゲーム法を用いた場合の負の互恵性に対する意図の効果. 日本社会心理学会第 53 回大会, つくば国際会議場.

【Awards】

平成 27 年度 神戸大学学生優秀学術表彰, 2016 年３月.

2015 年度 優秀学会発表賞 English Session 部門 日本グループ・ダイナミックス学会第 62 回大会, 2015 年 11 月.

2013 年度 大学院生海外学会発表支援制度 日本社会心理学会, 2013 年５月.