

# **Playing a Prisoner's Dilemma as an Assurance Game: Matrix Transformation and Production of Trust**

by

**Toshio Yamagishi\***

and

**Toko Kiyonari**

( Hokkaido University, Japan )

## **Introduction**

One of the most consistent findings in the experimental gaming research on prisoner's dilemmas and social dilemmas (i.e., the n-person version of the PD) is a correlation between player's own behavior and expectations of the partner's behavior (see Dawes, 1980; Edney & Harper, 1978; Kollock, in press; Komorita & Parks, 1995, 1996; Orbell & Dawes, 1981; Pruitt & Kimmel, 1977; Stroebe & Frey, 1982; van Lange, Liebrand, Messick & Wilke, 1992; Yamagishi, 1990a, 1995 for the general reviews of the literature; see Dawes, McTavish & Shaklee, 1977; Marwell & Ames, 1979; Sato & Yamagishi, 1984; Tyszka & Grzelak, 1976; Yamagishi & Sato, 1986 for the correlation). The correlation between the player's own behavior and expectations of the partner's behavior is known to exist even in one-shot games.

Such a correlation is striking especially given the fact that expectations of partner's behavior should have no theoretical role in one-shot games where defection is the dominant choice. Why do people care about other players' choices when the consequences of their own choice do not depend on the choices of the partner?

One possible answer to this question provided by Dawes and his associates (Dawes, 1989; Orbell & Dawes, 1993) is that the correlation does not mean that people care about their partner's behavior. Instead, according to Dawes and his associates, the correlation is a product of players who project their own behavior upon others. According to this "projection hypothesis," people are assumed to use their own behavior as a sample in inferring other people's behavior. Although the projection of one's own behavior onto the partner may explain some of the

---

\* Please address all correspondences to Toshio Yamagishi, Faculty of Letters, Hokkaido University, N10 W7 Kita-ku, Sapporo, Japan 060; Email [Toshio@letters.hokudai.ac.jp](mailto:Toshio@letters.hokudai.ac.jp); Fax +81-11-706-3066

correlation, there are experimental findings suggesting that the projection is not the only reason for the correlation. For example, Yamagishi & Sato (1986) found that the magnitude of the correlation was much stronger in the conjunctive structure than in the disjunctive structure. That is, when all members' cooperation was required for the provision of a public good, the correlation was strong ( $r = .83$ ); however, when only one member's cooperation was required for the provision of a public good, the correlation was much weaker ( $r = .35$ ). The difference cannot be explained by the projection hypothesis. Yamagishi & Sato argue that the difference in the magnitude of correlation cannot be explained unless the correlation is considered to be produced by the participants who adjust their behavior to the expectations of the partner's behavior. The expectation for the partner's behavior is much more important in the conjunctive structure (in which one's cooperation can be totally useless unless all other members are also cooperative) than in the disjunctive condition (in which a public good is produced by a cooperative player even when all other members do not cooperate at all). This and additional experimental findings to be presented below confirm that people do in fact care about other players' choices in one-shot social dilemmas. Why do they care?

The second answer to this question is that it is because people are in fact playing an assurance game rather than a prisoner's dilemma game when they play a PD game. This idea was originally developed for explaining cooperation in iterated prisoner's dilemma games. Since one's own defection invites the partner to retaliate in future in iterated games, cooperation is a more gainful choice than defection insofar as the partner cooperates (and, of course, with the proper

combination of the discount rate and matrix entries). The shadow of the future (Axelrod, 1984) thus transforms the "given matrix" (Kelley & Thibaut, 1978) of an iterated PD into an "effective matrix" of an assurance game. This is basically the account of the effectiveness of the *Tit-For-Tat* strategy (cf., Axelrod, 1984). It is also at the core of the goal/expectation theory proposed by Pruitt & Kimmel (1977). Pruitt and Kimmel, after reviewing over a thousand experimental gaming studies, proposed the theory, according to which cooperation in experimental games requires both mutual cooperation as a goal and expectations that others will be cooperative. They argue that many of the participants, having experienced the absurdity of mutual defection, come to adopt mutual cooperation as their goal, rather than pursuing the immediately better outcome in each iteration. And yet, this is not enough to make them behave cooperatively because many of them are afraid that their willingness to achieve mutual cooperation may be wasted or, worse, exploited by the partner. While realizing the importance of mutual cooperation to secure their own interest, they are hesitant to cooperate when the partner is not cooperating. They cooperate when and only when they are convinced that others will cooperate as well.

Both Axelrod and Pruitt & Kimmel regard that the long-term nature of iterated games makes players behave as if they are playing an assurance game in which cooperation is a better choice insofar as the partner cooperates. The transformation of an iterated PD game into an assurance game presented above, however, does not explain the transformation in one-shot games where there is no "shadow of the future." And yet, we often observe the correlation in one-shot games. Even in one-shot games, players seem to care

about the choices of other players. Why do they care? In this paper, I will first present research findings showing that players of one-shot PD games, at least a substantial proportion of them, do in fact play a PD game as an assurance game. Then, I will present my answer to the question of why those players play the PD game into the assurance game. And, finally, I will discuss some of the implications of playing a one-shot prisoner's dilemma game as an assurance game on production of trust.

### **Evidence showing that Players of a One-shot PD Game Are in Fact Playing an Assurance Game**

#### **Desirability of the Four Outcomes**

Direct evidence that players of a one-shot PD game do in fact play it as an assurance game comes from players themselves who indicate that they prefer the outcome of mutual cooperation over the outcome of unilateral defection. According to Kollock (1997), participants of a vignette-type experiment indicated that they prefer the outcome of mutual cooperation over that of unilateral defection. Similarly, participants of an experiment by Watabe et al. (1996) played a one-shot PD in which each had a choice of giving (cooperation) or not giving (defection) 500 yen (about \$5) to the partner; the money given to the partner was doubled in value by the experimenter and the partner received 1,000 yen (about \$10). They were asked in the post-experimental questionnaire, "How satisfactory would it be to you if both you and your partner gave 500 yen?" on a response scale that varied from 1 for "not at all" to 7 for "very much

so." The mean response to this question was significantly higher than the mean response to a similar question asking "How satisfactory would it be to you if your partner gave 500 yen and you did not?" (6.22 vs. 4.62,  $t(147) = 7.66$ ,  $p < .001$ ). On the other hand, participants judged that not cooperating would yield a more satisfactory result than cooperation when the partner failed to cooperate (3.16 vs. 1.82,  $t(147) = 8.68$ ,  $p < .001$ ). The pattern was replicated with American participants by Hayashi et al. (1997); the corresponding means among American participants were (6.01 vs. 5.05,  $t(165) = 4.64$ ,  $p < .001$ ) and (3.66 vs. 1.76,  $t(165) = 11.21$ ,  $p < .001$ ). Table 1 shows the average desirability scores of the four outcomes. The table also shows the proportions of the participants whose stated preferences match the incentive structure of the assurance game and that of the prisoner's dilemma game. Based on the questionnaire responses, forty-one percent of the participants of Watabe et al.'s experiments actually played the PD game as an assurance game. and only 18% of them played it as a PD game. Among the American participants of Hayashi et al.'s experiment, 30% play an assurance game and 28% played a PD game.

Table 1 Average desirability scores and preference rankings of the four outcomes in PD games, and the proportions of participants who played the PD as such and as an assurance game in six experiments

		Mean Desirability				Mean Preference Ranking				Proportions			
Source	n	CC	DC	CD	DD	CC	DC	CD	DD	Desirability		Ranking	
										PD	AG	PD	AG
Watabe et al.	148	6.22	4.62	1.82	3.16					.182	.405		
Hayashi et al.	167	6.01	5.05	1.76	3.66					.281	.299		
Terai, 1995	81	6.33	4.71	1.94	3.79	1.38	2.46	3.47	2.70	.222	.457	.037	.395
Y & Kiyonari	79	6.44	4.82	2.13	3.59	1.43	2.43	3.43	2.68	.215	.418	.304	.468
Y, K, K & K	40	6.18	4.78	2.50	4.08	1.23	2.58	3.49	2.67	.175	.425	.200	.575
Y & Kikuchi	90					1.37	2.32	3.60	2.69			.267	.511

Table 2 The relationship between own behavior and expectations among those who play PD game as if it were an assurance game and those who play it as a PD game.

Source	Type of players	n#	n of C	Average expectations for the partner's cooperation		t for the difference	Correlation
				Cooperators	Defectors		
Watabe et al. + Hayashi et al. combined*	AG	59	35	.60	.27	5.71, p<.0001	.60, p<.0001
	PD	53	4	.48	.30	1.88, p<.07	.25, p<.07
Yamagishi & Kiyonari, 1997**	AG	17	13	.60	.48	1.49, ns.	.36, ns.
	PD	10	0	None of the participants cooperated.			
Yamagishi et al., 1997	AG	23	16	.54	.30	3.06, p<.01	.55, p<.01
	PD	8	3	.53	.46	0.06, ns.	.23, ns.
Yamagishi & Kikuchi, 1997	AG	46	23	.80	.32	6.45, p<.0001	.70, p<.0001
	PD	24	1	Only one participants cooperated.			
Terai, 1995	AG	36	19	.63	.31	4.90, p<.0001	.64, p<.0001
	PD	18	1	Only one participants cooperated.			

\* three, logically identical conditions

\*\* in a condition in which participants were asked of the expectations

# The data include only those for whom the stated preferences strictly confirm the PD or assurance game structure.

In three other experiments of one-shot PD games conducted by Yamagishi and his associates (Terai, 1995; Yamagishi, Kikuchi, Kiyonari & Kosugi, 1997; Yamagishi & Kiyonari, 1997), participants expressed their preferences for the four outcomes by assigning preference orders to the four outcomes in addition to

evaluating the desirability of each outcome separately. The average of those preference rankings were consistent with the average desirability scores of the four outcomes, indicating that on average participants of those experiments played the PD game given by the experimenter as an assurance game. The proportion of the

participants who played the game as an assurance game (hereafter called the “**AG players**”) ranged from 40% to 58%, and the proportion of the participants who played it as a PD game (hereafter called the “**PD players**”) ranged from 4% to 30% (see Table 1). Finally, participants of the sixth experiment (Yamagishi & Kikuchi, 1997) shown in Table 1 expressed their preferences for the four outcomes only by ranking the desirability of the four outcomes. The results of this experiment were also consistent with those from the previous experiments, showing that on average participants played the given game of a prisoner’s dilemma as an assurance game.

The above findings provide evidence that the correlation between one’s own behavior and expectations of the partner’s behavior is produced by the fact that some players of a prisoner’s dilemma game play the PD game as an assurance game. Further evidence of this explanation of the behavior-expectation correlation in one-shot PD games comes from the finding that the **correlation exists only among those who play the game as an assurance game**. This finding comes from a re-analysis of Watabe et al.’s (1996) and Hayashi et al.’s (1997) experimental data. Their experiments included six conditions. Two of them have been presented so far. One of the remaining four conditions was the ordinary one-shot PD condition in which players made decisions simultaneously. The other conditions involved sequential PD games. In the self-first/no-knowledge condition, the participant made a decision first knowing that his/her decision will not be revealed to the partner before the partner made his/her decision. In the other-first/no-knowledge condition, the participant made a decision after they were informed that the partner had already made a decision, though they

were not informed of the content of the partner’s decision. Logically, these three conditions—simultaneous condition, self-first/no-knowledge condition, other-first/no-knowledge condition—are identical, and thus I pooled the participants in those three conditions together in the following analysis.<sup>1</sup> Among the participants in those three conditions, expectations of the partner’s choice were strongly related to their own choice ( $r = .60, p < .0001$ ; the mean expectation of the partner’s cooperation among cooperators = .56, among defectors = .27,  $t(57) = 5.71, p < .0001$ ). On the other hand, the correlation was much weaker among the participants who played the game as a PD game ( $r = .25, p < .07$ ;  $m$  among cooperators = .48,  $m$  among defectors = .30,  $t(51) = 1.88, p < .07$ ).

It should be noted, however, that only very few (4 of 53) of the PD players cooperated in this experiment. And as shown in Table 2, the number of cooperative PD players is very small in the other experiments as well. Thus, the weak behavior-expectation correlation observed among the PD players may not be reliable, and the interpretation of the above finding as suggesting the role of matrix transformation as the cause of the behavior-expectation correlation requires caution. However, results of another experiment by Jin & Shinotsuka (1996) provided a reassurance for that interpretation. The participants’ choices in this experiment were continuous rather than dichotomous. They decided how much of the endowment of 200 yen to give to the partner, and the partner received twice the amount provided by the

---

<sup>1</sup> The remaining condition was the self-first/knowledge condition in which the participant made a decision first knowing that his/her decision would be revealed to the partner before the partner

participant.<sup>2</sup> Thus, we do not face the above problem caused by the paucity of cooperative PD players. The amount participant decided to give to the partner in this experiment was regressed on the expectations for the partner's provision amount together with a dummy variable for the AG players versus the PD players and the interaction term. The effect of the expectation on the participant's own behavior was significantly stronger among the AG players (unstandardized regression coefficient,  $b = 1.22$ ) than among the PD players ( $b = .50$ ),  $t(46)$  for the interaction effect = 2.62,  $p < .05$ . The relationship between one's own behavior and expectations in this study was thus shown to be much more pronounced among the AG players than among the PD players. These results, taken together, strongly suggest that the behavior-expectation correlation is mostly produced by those who play a PD game as an assurance game.

### **Behavior in Sequential PD Games**

The preferences for the four outcomes stated by the participants of the above experiments consistently indicate that more participants play the PD game as an assurance game than as such. However, it is possible that the stated preferences are just "lip service" and do not reveal their true preferences. Experiments by Watabe et al. (1996) and Hayashi et al. (1997) provide a safeguard against this criticism, however, and indicate that the participants' stated preferences for the four outcomes are not mere "lip service." Those experiments

included two conditions in which participants were first informed of the partner's choices and then decided whether to cooperate or defect. In one of these conditions, participants were informed that the partner had already made a decision to cooperate. In the other condition, they were informed that the partner had already made a decision to defect. The overwhelming majority of the participants (20 of the 23 Japanese participants in Watabe et al.'s experiment, and all of the 13 American participants in Hayashi et al.'s experiment) in the second condition defected, as everyone would expect. How about the first condition? If the participants' actual preferences were consistent with the monetary values assigned to the four outcomes, they would have defected in the first condition as well. Quite contrary to this expectation, the majority of the participants in this condition who knew that the partner had already made a decision to cooperate (15 of the 20 Japanese participants in Watabe et al.'s experiment, and 11 of the 18 American participants in Hayashi et al.'s experiment) cooperated. Furthermore, practically all (9 of the 10 Japanese and all of the 6 American) participants in this condition who expressed the preference pattern of the assurance game (the AG players) cooperated. In contrast, the overwhelming majority (two of the three Japanese and all of the three American participants) in this condition who expressed the preference pattern of the prisoner's dilemma (the PD players) defected. These results show (1) that many of the participants play the prisoner's dilemma game as an assurance game, and (2) that their actual behavior is consistent with their stated preferences.

---

made his/her decision.

<sup>2</sup> Only one of three within-subjects conditions, in which the participant played a PD game with an unidentified partner, was used for this analysis. In the other two conditions, the participant played a PD game with either an ingroup member or an outgroup member.

## **Utility Transformation and**

## Matrix Transformation

The research findings presented above seem conclusive. We now can conclude with a fair amount of confidence that there are people who play a one-shot PD game as an assurance game. In this section, I will address the question of how this happens. More specifically, I compare two possible accounts of this phenomenon. One possibility is that those people accommodate the partner's outcomes into the overall evaluation of the desirability of the outcomes. That is, one possible account for the AG players is that they derive utilities from the outcomes the partner receives as well as from the outcomes they themselves receive. Let me call this the **utility transformation** approach. We may also call this the motivational approach. The second possibility is that people have a tendency to perceive a PD game structure, or almost any form of the interdependence structure, as an assurance game. They are not deriving utilities directly from the outcomes the partner receives; instead, they perceive that it is a better choice for them to cooperate insofar as the partner cooperates as well. According to this **matrix transformation** approach, or the cognitive approach, the AG players are the ones who have a default assumption about the nature of the interdependent relation and use the assumption in most decision making situations involving an interdependent relation unless salient information about the nature of the relation is provided. For them, perceiving a PD game as an assurance game is a heuristics, or a ready-made routine for processing cognitive information. The experimental findings shown below are more consistent with the matrix transformation view rather than with the utility transformation view.

## Social Value Orientation

The standard way that has been used to interpret individual differences in experimental gaming adopts "social motivation" or "social value orientation" as the key concept. Social motivation or social value orientation refers to the relative importance a player assigns to the outcomes to him/herself and to the partner. Starting with the use of the decomposed game technique (Messick & McClintock, 1968), numerous studies have explored, using various measurement techniques, the relationship between social motivation and behavior in experimental games (e.g., Kuhlman, Camac & Cunha, 1986; Kuhlman & Marshello, 1975; Kuhlman & Wimberley, 1976; Liebrand, 1984, 1986; Liebrand, Wilke, Vogel & Wolters, 1986; McClintock & Liebrand, 1988). According to this general approach, the difference between cooperators and defectors are sought in the importance they assign to (or utilities they derive from) the partner's outcomes in comparison to their own outcomes. "Cooperators" cooperate even in a one-shot PD game since for them the outcome the partner receives is as important (i.e., they derive as much utilities from) as the outcome they themselves receive. "Individualists" defects in one-shot PD games since they care only about their own outcomes. "Competitors" are the ones who enjoy the relative advantage they have over the partner; they derive a positive utilities from seeing the partner to lose even at a cost of a smaller loss to themselves. They naturally defect in one-shot PD games.

The motivational approach (or utility transformation approach) represented by the social value orientation literature, however, faces a difficulty in explaining the aforementioned, tremendous difference in the cooperation rate observed in the sequential PD experiments by Watabe et al.

and by Hayashi et al. In those experiments, practically no one cooperated when the players knew that the partner had already defected. In contrast, majority of the players cooperated when they knew that the partner had already chosen to cooperate. If the reason for cooperation in the prisoner's dilemma game is based on the utilities players derive from the outcomes to the partner (i.e., "I'm happy because my partner is happy"), then, what the partner does should not matter. The utility transformation approach cannot explain the player's behavior that varies depending on the partner's behavior.

Facing this difficulty, an alternative, a more cognitive interpretation of the social value orientation has been proposed. According to this alternative view of social value orientation (Kuhlman et al., 1986), the stable differences found among different groups of people (cooperators, individualists, competitors, etc.) are "due to differences in beliefs as to the best way to maximize one's own welfare rather than to differences in preferences as to what to maximize" (p.164). According to this view, consistent defectors in prisoner's dilemmas and social dilemmas are the ones who consider mutual cooperation a "highly desirable, but impossible dream." Using our terms, they are playing an assurance game while expecting that others are not willing to cooperate.

The above interpretation of the individual differences traditionally conceptualized as ones reflecting their social value orientation can thus be re-conceptualized from the matrix transformation point of view. Instead of directly deriving utilities from the outcomes given to the partner, the AG players perceive the prisoner's dilemma structure as that of an assurance game. Therefore, the cooperate when and only when the partner is expected to cooperate as well.

This interpretation of social value orientation as the subjective game type is also consistent with the classic finding in the experimental gaming research by Kelley and Stahelsky (1970). In general, people tend to perceive others as similar to themselves. For example, cooperators tend to see most people as cooperative, competitors tend to see most others as competitive, and individualists tend to see most others as individualistic (Kuhlman and Wimberley, 1976). However, Kelley & Stahelsky found this to be true only for defectors. They found that defectors always think that other people are not cooperative. On the other hand, cooperators tend to think that some people are cooperative and others are not, and adjust their choices to their expectations of the partner's behavior. According to the interpretation that "cooperators" are the ones who plays a PD game as an assurance game, they should be the ones who are paying attention to whether the partner would cooperate or defect rather than simply project their behavior onto the partner. In contrast, the projection of own behavior onto the partner may be a common phenomenon among the PD players since their behavior is consistent. In sum, either the variability in the player's behavior reflecting the behavior of the partner found by Watabe et al. or the variability in the sensitivity between cooperators and defectors found by Kelley & Stahelsky cannot be explained by the utility transformation approach. To explain those variabilities require the matrix transformation approach.

### **Cooperation with Ingroup and Outgroup Members**

Another area of research that provide a good ground for testing between the two approaches—i.e., the matrix transformation

and the utility transformation approaches—is cooperation with ingroup and outgroup members. Players of experimental games have been known to cooperate more with a member of the same group (ingroup member) than with a member of another group (outgroup member). This effect of group identity is known to exist even when the groups are the “minimal groups” (Brewer & Kramer, 1986; Kramer & Brewer, 1984; Wit & Wilke, 1992). “Minimal group” is a term used by Tajfel and his associates (Tajfel, 1982; Tajfel, Billig, Bundy & Flament, 1971; Turner, 1987) to describe the grouping of people devoid of history or any kind of real interactions.<sup>3</sup> In an epoch making experiment, Tajfel et al. (1971) found that participants practice ingroup favoritism even in the minimal group situation, allocating more money to a member of their own “minimal group” who shared only a trivial category with them and less to a member of the other group. The explanation they gave to this effect of group identity is, roughly translated into our terms, that participants have a “competitive” value orientation. They derive positive utilities from seeing others who share the same social category with them to prosper relative to the others. According to their accounts, the group identity effect on ingroup favoritism in reward allocation is based on the utilities they derive from the outcomes given to members of their own “ingroup” (i.e., those who share the same social category with them). In other words, their accounts of the group identity effect is based on utility transformation.

Jin, Yamagishi and their associates (Karp, Jin, Yamagishi & Shinotsuka, 1993;

Jin & Yamagishi, 1997; Jin, Yamagishi & Kiyonari, 1996) argue that the group identity effect in the minimal group situation can better be explained based on the “group heuristics” or “group cooperation heuristics.” That is, they argue that people practice ingroup favoritism even in the minimal groups because they expect a similar, favorable responses from ingroup members, and only from ingroup members. When people use the group heuristics, they assume that their behavior is a part of a generalized exchange system. The group heuristics is an assumption invoked when people face decisions in an interdependent situation, an assumption that the situation is a part of a larger generalized exchange system. Group-related cues work as primes to activate the heuristics, and thus participants of minimal group experiments behave toward ingroup members as if they are expecting generalized reciprocity (i.e., the expectation that the favor they give to an ingroup member will somehow be returned, not necessarily from the same person but possibly from someone in the same group). They demonstrated through a series of experiments that ingroup favoritism in the minimal group experiments is a product of the group heuristic and not the product of utility transformation.

Jin & Shinotsuka (1996) and Jin & Yamagishi (1997) applied this idea of group heuristics in explaining the higher level of cooperation with ingroup than outgroup members. The basic idea behind those experiments is that matrix transformation from the PD game to the assurance game is a part of this group heuristics. Given the existence of a generalized exchange system, playing a PD as an assurance game is a more gainful strategy than playing a PD as such since the favor given to a member is eventually returned. When a system of

---

<sup>3</sup> In this sense, one-shot prisoner's dilemma games between anonymous players is a minimal group. There is no real interdependency in the sense that players can affect each other's behavior.

generalized exchange operates in a group, whether each relation involves a prisoner's dilemma structure or an assurance game structure does not matter; cooperation will be rewarded by cooperation from the group members and defection will be punished by defection or, worse, by expulsion from the group. Making decisions based on the group heuristics—and thus treating a PD as an assurance—is expected to have an adaptive advantage when it is not certain whether or not a particular decision scene is a part of a larger generalized exchange system. In such a situation, players can make two types of errors. They make Type I error when they erroneously conclude that the null hypothesis (non existence of a generalized exchange system) is wrong when it in fact is true. When they make this type of error, they forgo extra gains they could have obtained by treating the situation as a true one-shot PD. On the other hand, they make Type II error when they erroneously conclude that the null hypothesis is true; that is, they conclude that a generalized exchange system does not exist when it in fact does exist. The consequences of making Type II error can be far more serious than those of making Type I error. The group heuristics protects people from making Type II error in an ambiguous situation, although it increases the chance of making Type I error.

Participants of Jin & Shinotsuka's (1996) experiment were first categorized into two groups ostensibly based on the result of a "perception experiment" (and actually randomly). In the perception experiment, they estimated the angles of black fans displayed on the white background and white fans on the black background. Then, they were divided into two categories, the "black perceivers" who are supposed to overestimate the angles of black fans and the overestimators of the angles of white fans (white perceivers).

After they had completed this perception experiment and had been classified either as a black perceiver or a white perceiver, the participants played the same one-shot PD game three times, each with a different partner; i.e., once with a member of the same group, once with a member of the other group, and once with a partner whose group identity is not known. Only the group identity, not the individual identity, of the partner was revealed to the participant in the first two conditions. The PD game was constructed in such a way that each player decided how much of an endowment of 200 yen to give to the partner. The partner received twice the amount given by the player. The result of this experiment successfully replicated the previous findings concerning the effect of group identity. Participants gave an average of 99.20 yen to an ingroup member and 80.34 yen to an outgroup member. When the group identity of the partner was not known, the amount they gave to the partner was 91.70 yen, a halfway between the ingroup and the outgroup conditions. The effect of the partner's group identity was statistically significant,  $F(2, 174) = 5.63, p < .01$ . The partner's group identity also had a significant, and even stronger effect on the amount participants expected to be given by the partner; participants expected that an ingroup member, an outgroup member, and unidentified member would give an average of 100.45 yen, 74.02 yen, and 87.24 yen, respectively,  $F(2, 174) = 12.33, p < .0001$ . The fact that group identity had a greater effect on expectations than on actual behavior was certainly inconsistent with the projection hypothesis of the behavior-expectation correlation and was consistent with the view that participants adjusted their behavior to their expectations of the partner's behavior. Furthermore, the effect of group identity on

the behavior (i.e., the amount the participant gave to the partner) disappears when expectations are statistically controlled,  $F(2, 172) = 2.04$ , ns., whereas the effect of group identity on expectations does not disappear even when the behavior is controlled,  $F(2, 172) = 7.93$ ,  $p < .0001$ . These results strongly suggest that the effect of group identity on actual behavior was mediated by its effect on expectations rather than the other way around.<sup>4</sup>

Furthermore, participants filled out a post-experimental questionnaire that included a "perception of interdependence scale." The scale consists of items such as "if you cooperate with others, others will cooperate with you," "if you want to do well, you better help others and being helped by others," "helping others will eventually help oneself," "helping others will not do any good to oneself" (reverse item), and so on; it is supposed to measure the belief that one's *personal* success requires cooperation with others. It should be noted that the scale is not intended to measure moral belief for the importance of mutual cooperation; rather, it is intended to measure belief for the importance of mutual cooperation *as a means* to achieve personal success. The participants' responses to this scale were significantly related to the type of games they played; the mean scale value for the AG players and PD players were 4.11 and 3.77 on a 7-point scale, respectively, and the difference was significant,  $t(48) = 2.01$ ,  $p < .05$ . When participants were classified into two groups based on the medium score, the effect of the partner's group identity disappeared among those who scored low on this scale. Among those who

scored high on this scale (high interdependants), the average amount they gave to the ingroup, unidentified, and outgroup partner were 120.91 yen, 104.77 yen, and 91.82 yen, respectively. On the other hand, among the low interdependants, those averaged were 77.50 yen, 78.64 yen, and 68.86 yen. High interdependants not only gave more to the partner; they favored ingroup members more strongly than did low interdependants. Those results suggest that AG players are those who consider mutual cooperation as an effective means to achieve their goals

The possibility suggested in the above study that the effect of group identity on actual behavior in one-shot prisoner's dilemmas is mediated by expectations of the partner's behavior was further investigated by Jin and Yamagishi (1997). Participants of their experiment played the same prisoner's dilemma game five times, once in each of the five within-subjects conditions. Two of the five experimental conditions are the standard ingroup/outgroup conditions. The group identity was manipulated through a "separate experiment" on "perception" conducted prior to the prisoner's dilemma experiment, in which the participants rated several times their preferences for two paintings—a Klee's painting and a Kandenski's painting—displayed on a screen for 5 seconds. Participants were then classified into two groups, the Kandenski group and the Klee group supposedly based on their preferences (but in fact randomly). The third condition was a control condition in which the partner's group identity was not known. These three conditions constituted a replication of the above experiment and other, similar experiments. The unique feature of this experiment was in the additional of the following two extra conditions. In the fourth condition,

---

<sup>4</sup> See Yamagishi (1990b) for a similar conclusion concerning the effect of group size on actual behavior and expectations.

participants played a prisoner's dilemma game with an ingroup member, but they were told that the partner would not know the group identity of the participant. That is, the participant knew that the partner was an ingroup member, but the partner did not know the group identity of the participant. Similarly, participants in the fifth condition played a prisoner's dilemma game with an outgroup member who did not know the group identity of the participant. Thus, except for the control condition in which participants did not know the group identity of the partner, group identity of the partner was crossed with the knowledge the partner had about the participant's group identity. In each of the five prisoner's dilemma games each participant was given an endowment of 100 yen and was asked how much of it to give to the partner (in increments of 10 yen). The partner received twice the amount given by the participant.

According to the utility transformation view of the effect of group identity, the knowledge the partner has of the participant's group identity should have no effect. In contrast, according to the group heuristics point of view, the knowledge should be critical for the effect of group identity to occur. According to the group heuristics perspective, cooperation in a one-shot PD is mostly produced by those who assume, by default, that their decision constitutes a part of a generalized exchange system. For them, the PD game they play is a part of a larger generalized exchange system. They thus cooperate insofar as the partner is expected to cooperate as well. And the expectation of the partner's cooperation comes from the expectation that the partner would act cooperatively toward his/her own group members. In the no knowledge condition, however, the partner does not know that the participant is an ingroup (or an

outgroup) member, and thus even the partner of the same group is not expected to cooperate with the participant whose group identity is unknown to him/her. The results of the experiment shown in Table 3 is consistent with this group heuristic point of view, and is inconsistent with the utility transformation view that predicts only the main effect of the partner's group identity. The partner's group identity affected the subject's cooperation level greatly (30.69 versus 20.50) when the partner knew of the subject's group identity but not when the partner did not have that knowledge (23.69 versus 19.04). The former difference was significant, and the latter was not. Furthermore, a set of planned comparisons indicates that the first cell (in which participants played with an ingroup member who knew of their group identity) was significantly different from the other cells, and the other four cells did not significantly differ from each other.

The results from those two experiments consistently indicate that the effect of the partner's group identity on the player's behavior is mostly mediated by its effect on expectations of the partner's behavior. For expectations of the partner's behavior to play such a role, the player should be playing an assurance game rather than a prisoner's dilemma game. These findings add to the evidence that the matrix transformation is a matter of cognition about the nature of the interdependence, and not so much a matter of utilities or motivations .

Table 3 Average amount subjects contributed in Jin & Yamagishi's (1997) experiment

Condition		Average contribution
Partner's Group Identity	Partner's Knowledge	
Ingroup	Yes	30.69 (sd = 23.02)
Ingroup	No	23.69 (sd = 20.64)
Outgroup	Yes	20.50 (sd = 21.90)
Outgroup	No	19.04 (sd = 18.00)
Control (partner's identity unknown)		20.74 (sd = 20.97)

### Production of Trust

I started this paper with the behavior-expectation correlation and concluded that the correlation is produced by the AG players who adjust their behavior to the partner's expected behavior. The AG players are concerned with whether or not they can trust the partner. In this section, let me call them "consumers" of trust since they are concerned with how much trust is being provided by others. We are all consumers of trust in this sense in many occasions. As a consumer of trust, we are concerned with the trustworthiness of our potential interaction partners in the same manner as consumers of automobiles are concerned with the quality of their potential purchase. At the same time, we often are "producers" of trust as well, trying to make ourselves trusted by the partner. The most straightforward way of producing trust is to post a hostage; one can make oneself trusted by posting a hostage. Posting a hostage produces a trust, making him/herself trusted by the partner, and thus obtain the outcome that he/she could otherwise not have obtained. For example, a used car dealer can make him/herself trusted by providing a guarantee for the car he/she sells, and thus makes it easier to sell cars to customers. Understanding the advantage of producing trust, however, it is

sometimes difficult to post a proper hostage. One reason for this difficulty, of course, is the unavailability of a proper hostage. Another possible reason is the fear that the hostage may be misused.

In an unpublished experiment, Yamagishi & Kiyonari (1997) addressed the issue of how this fear prevents production of trust. They gave the participants a choice between two types of games in addition to the ordinary choice between cooperation and defection. One type of game is an ordinary one-shot PD. Each player decided whether or not to give an endowment of 500 yen to the partner; when a player gave 500 yen, the partner received twice the amount or 1,000 yen. The second type is a sequential version of the same one-shot PD in which the participant was to play first. The partner would then decide between C and D with the full knowledge of the participant's choice.<sup>5</sup> This is a situation in which production of trust is highly effective especially for the nervous AG players to achieve their desired goal. The nervous AG players are those who play a PD game as an assurance game while expecting that the partner would not cooperate. There are at least two possible reasons for that

<sup>5</sup> There was no partner in this experiment. That is, all participants play the role of the first player when they chose the sequential game.

expectation. One reason is that the partner is expected to be greedy, or that the partner is playing the PD game as such. Facing this possibility, the AG players have no alternatives than defecting in self-defense. This is not, however, the case when the reason for the partner's defection is expected to be based on the perceived fear of the partner. That is, the partner may be expecting that he/she is playing against a defector in the same way as the AG player is concerned about the partner's defection. The AG players who are willing to cooperate otherwise may not be able to do so **when they are afraid that the partner is also afraid**. A key to success in such a situation for the nervous AG players is the production of trust—letting the partner to trust him/her. Volunteering to reveal one's choice of cooperation to the partner in a sequential PD is a very effective way of **producing** the needed **trust** in this situation.

Logically, there are only two rational combinations of choices for the participant of this experiment—i.e., (1) to defect and not to reveal and (2) to cooperate and to reveal. The other combinations, to defect and reveal and to cooperate and not to reveal, are not rational. First, if one cooperates, revealing the choice to the partner before the partner makes his/her decision only increases the chance that the partner cooperates. Some of the partners may not be affected by the choice revealed to them. On the other hand, there is no logical ground to expect that the partner who is willing to cooperate in the simultaneous game changes his/her mind and defects as he/she is revealed the cooperative choice of the first player. If the partner is a PD player, he/she will defect anyway regardless of the choice of the first player; revealing the cooperative choice to the partner then will have no effect on the partner's choice. If the partner is an AG

player, especially a nervous AG player, revealing the cooperative choice by the first player provides an assurance to the partner that he/she needed to make him/her act cooperatively as well. Thus, if one cooperates one should reveal the choice to the partner. Secondly, if one defects, revealing his/her defection to the partner would make the possibility of the partner's cooperation practically zero; thus, it is better not to reveal own defection. Thus, it is expected that cooperators in this experiment would reveal their choices (i.e., choose the sequential game) and that defectors would not reveal their choices (i.e., choose the simultaneous game). Kiyonari & Yamagishi (1997) examined these logical predictions, and found that their participants did not act so logically. While most of the defectors (36 of the 39) chose the simultaneous game rather than the sequential game as predicted, over 40 percent (17 of the 40) of the cooperators failed to choose the predicted sequential game. They were given a chance to produce trust by volunteering to reveal their choice to the partner, and failed to take advantage of this chance. Why did they fail? Why did they choose not to reveal their choice of cooperation to the partner?

In order to answer this question, we compared two types of cooperators—those who chose the sequential game and those who chose the simultaneous game—on their post-experimental questionnaire responses. The answer suggested from the post-experimental questionnaire data is that the second type of cooperators who chose the simultaneous game had an unfounded fear that revealing their cooperative choice to the partner put themselves in a vulnerable position, leaves them at the mercy of the partner. Specifically, those who chose the simultaneous game agreed more strongly

with the following items than those who chose the sequential game: "I thought that the partner would take advantage of it if I reveal my choice to him/her" (4.55 on a 7 point scale versus 3.38,  $t(20) = 1.49$ ,  $p < .08$ , one-tailed); "It is not fair that some people are revealed of the partner's choice and some are not" (3.33 vs. 1.92,  $t(20) = 2.17$ ,  $p < .05$ ); "I felt uneasy for revealing the choice since I would then be at the mercy of the partner" (3.67 vs. 2.38,  $t(20) = 2.01$ ,  $p < .05$ , one-tailed); "I was wondering if the partner might give in to the temptation for not giving the 500 yen if I revealed my choice" (5.00 vs. 4.08,  $t(20) = 1.53$ ,  $p < .08$ , one-tailed); "I did not like to reveal my choice to the partner since it made my position unequal to that of the partner" (4.22 vs. 1.46,  $t(10) = 6.04$ ,  $p < .0001$ ).<sup>6</sup> The correlation between the type of game they chose (a dummy variable with the choice of the simultaneous game as coded as one) and the response to those questions were generally high;  $r = .32, .44, .41, .32$ , and  $.84$ , in the order of the questions presented above. The AG players who were willing to cooperate in a one-shot PD game are thus shown to fail to produce the needed trust when they felt that revealing their choice put them in a disadvantaged position, leaving them at the mercy of the partner.

As mentioned earlier, there is no rational basis to expect that anyone who cooperates in the simultaneous game change their mind and defects when he/she finds out that the partner has decided to cooperate. Furthermore, the data from Watabe et al.'s and Hayashi et al.'s experiments showing that more people cooperated when they were informed of the partner's cooperative

choice than in the simultaneous game provide empirical basis for this logical conclusion. Thus, the fear that revealing one's willingness to cooperate makes him/herself vulnerable is unfounded. And yet, this fear seems to be the major obstacle for the production of trust. What is the nature of this fear, and why some people are haunted with this unfounded fear? One possible explanation for this fear is that the group heuristics is activated only when mutual dependency exists. The group heuristics is a default expectation of generalized exchange such that a favor one provides to someone will eventually be paid back. For the second player whose partner has already made a decision, it may be difficult to feel that his/her behavior has any effect on his/her outcome through its effect on other peoples' behavior. The fear can be based on the expectation that the partner does not treat the relation as a part of a generalized exchange system. Since we do not have empirical data on this issue, this and other possible explanations of this unfounded fear needs to be empirically examined in future research. Meanwhile, we may conclude that whatever the cause of the fear is, production of trust can be prevented because of this unfounded fear. Turning consumers of trust into producers of trust requires this unfounded distrust of the partner be removed somehow. Transforming PD games into AG games produces needs for trust, but it is not enough by itself for the production of trust.

## References

- Axelrod, Robert. 1984. *The evolution of cooperation*. New York: Basic Books.
- Brewer, Marilyn B., & Kramer, Roderick M. 1986. Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing. *Journal of Personality and*

---

<sup>6</sup> This analysis was conducted only with cooperators. Furthermore, for a technical reason only a half of the participants were asked of this question.

- Social Psychology* 50, 543-549.
- Dawes, Robyn M. 1980. Social dilemmas. *Annual Review of Psychology* 31, 169-193.
- Dawes, Robyn M. 1989. Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology* 25, 1-17.
- Dawes, Robyn M., McTavish, Jeanne, & Shaklee, Harriet. 1977. Behavior, communication and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology* 35, 1-11.
- Edney Julian J., Harper Christopher S. 1978. The commons dilemma: A review of contributions from psychology. *Environmental Management* 2, 491-507.
- Hayashi, Nahoko, Ostrom, Elinor, Walker, James, & Yamagishi, Toshio. 1977. Reciprocity, trust, and the illusion of control. Mimeo: Workshop in Political Theory and Policy Analysis, Indiana University, Bloomington.
- Jin, Nobuhito, & Yamagishi, Toshio. 1997. Group heuristics in social dilemmas. *Japanese Journal of Social Psychology* 12, 190-198. (In Japanese with an English abstract)
- Jin, Nobuhito, Yamagishi, Toshio, & Kiyonari, Toko. 1996. Bilateral dependency and the minimal group paradigm. *The Japanese Journal of Psychology* 67, 77-85. (In Japanese with an English abstract)
- Karp, David, Jin, Nobuhito, Yamagishi, Toshio, & Shinotsuka, Hiromi. 1993. Raising the minimum in the minimal group paradigm. *The Japanese Journal of Experimental Social Psychology* 32, 231-240.
- Kelley, Harold H., & Stahelsky, A. J. 1970. The social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology* 16, 66-91.
- Kelley, Harold H., & Thibaut, J. W. 1978. *Interpersonal relations*. New York: Wiley.
- Kollock Peter. 1997. Transforming social dilemmas: Group identity and cooperation. In *Modeling Rational and Moral Agents*, ed. P. Danielson, pp. 186-210. Oxford: Oxford University Press
- Kollock, Peter. In Press. Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*.
- Komorita, Samuel S., & Parks Craig D. 1995. Interpersonal relations: mixed-motive interaction. *Annual Review of Psychology* 46, 183-207.
- Komorita, Samuel S., & Parks Craig D. 1996. *Social Dilemmas*. Westview Press.
- Kramer, Roderick M., & Brewer, Marilyn B. 1986. Social group identity and the emergence of cooperation in resource conservation dilemmas. Pp.205-234 in Henk A. M. Wilke, David M. Messick and Christel G. Rutte (eds.), *Experimental social dilemmas*. Frankfurt am Main: Verlag Peter Lang.
- Kuhlman, David. M., Camac, C. R., & Cunha, D. A. 1986. Individual differences in social orientation. Pp.151-176 in Henk A. M. Wilke, Dave M. Messick and Christel G. Rutte (eds.), *Experimental social dilemmas*. Frankfurt am Main: Verlage Peter Lang.
- Kuhlman, David M., & Marshello, A. F. J.. 1975. Expectations of choice behavior held by cooperators, competitors, and individualists across four class of experimental game. *Journal of Personality and Social Psychology* 32, 922-931.
- Kuhlman, David. M., & Wimberley, D. L. 1976. Expectations of choice behavior held by cooperators, competitors, and individualists across four classes of experimental games. *Journal of Personality and Social Psychology* 34, 69-81.
- Liebrand, Wim B. G. 1984. The effect of social motives, communication and group size on behavior in an n-person multi-stage mixed-motive game. *European Journal of Social Psychology* 14, 239-64.
- Liebrand, Wim B. G. 1986. The ubiquity of social values in social dilemmas. Pp. 113-133 in H. A. M. Wilke, D. M. Messick and C. G. Rutte (eds.), *Experimental social dilemmas*. Frankfurt am Main: Verlag Peter Lang.
- Liebrand, Wim B. G., Wilke, Henk A. M., Vogel, R., & Wolters, F. J. M. 1986. Value orientation and conformity in three types of social dilemma games. *Journal of Conflict Resolution* 30, 77-97.
- Marwell, Gerald, & Ames, Ruth E. 1979.

- Experiments on the provision of public goods, I: Resources, interest, group size, and the free-rider problem. *American Journal of Sociology* 84, 1335-1360.
- McClintock, Charles G., Liebrand, Wim B. G. 1988. Role of interdependence structure, individual value orientation, and another's strategy in social decision making: A transformational analysis. *Journal of Personality and Social Psychology* 55, 396-409.
- Messick, David M., & McClintock, Charles G. 1968. Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology* 4, 1-25.
- Orbell John., & Dawes, Robyn M. 1981. Social dilemmas. In *Progress in applied social Psychology*, Vol. 1, ed. G. M. Stephenson & J. M. Davis, pp. 37-65. New York: Wiley & Sons
- Orbell, John, & Dawes, Robyn M. 1993. Social welfare, cooperators' advantage, and the option of not playing the game. *American Sociological Review* 58, 787-800.
- Pruitt, Dean G., & Kimmel, Melvin J. 1977. Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Review of Psychology* 28, 363-392.
- Sato, Kaori, Yamagishi, Toshio. 1984. Two psychological factors in the problem of public goods. *Japanese Journal of Experimental Social Psychology* 26, 89-95. (In Japanese)
- Stroebe, Wolfgang, Frey, Bruno S. 1982. Self-interest and collective action: The economics and psychology of public goods. *British Journal of Social Psychology* 21, 121-137.
- Tajfel, H. 1982. Social psychology of intergroup discrimination. *Scientific American* 223, 96-102.
- Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. 1971. Social categorization in intergroup behaviour. *European Journal of Social Psychology* 1, 149-178.
- Terai, Shigeru. 1995. Cooperation in a prisoner's dilemma and the perception of interdependence. Master's Thesis, Faculty of Letters, Hokkaido University, Sapporo, Japan. (In Japanese)
- Turner, J. C. 1987. *Rediscovering the social group*. Basil Blackwell.
- Tyszka, Tadeusz, Grzelak, Janusz L. 1976. Criteria of choice in non-constant zero-sum games. *Journal of Conflict Resolution* 20, 357-376.
- van Lange, Paul A. M., Liebrand Wim B. G., Messick, David M., & Wilke, Henk A. M. 1992. Social dilemmas: The state of the art—introduction and literature review. In *Social dilemmas: Theoretical issues and research findings*, ed. W. B. G. Liebrand, D. M. Messick, & H. A. M. Wilke, pp. 3-28. Oxford: Pergamon Press.
- Watabe, Motoki, Terai, Shigeru, Hayashi, Nahoko, & Yamagishi, Toshio. 1996. Cooperation in the one-shot prisoner's dilemma based on expectations of reciprocity. *Japanese Journal of Experimental Social Psychology* 36, 183-196. (In Japanese with an English abstract)
- Wit, Alphons P., & Wilke, Henk A. 1992. The effect of social categorization on cooperation in three types of social dilemmas. *Journal of Economic Psychology* 13, 135-151.
- Yamagishi, Toshio. 1990a. *Social dilemmas*. Tokyo: Science Press. (In Japanese)
- Yamagishi, Toshio. 1990b. Factors mediating residual effects of group size in social dilemmas. *The Japanese Journal of Psychology* 61, 162-169. (In Japanese)
- Yamagishi, Toshio. 1995. Social dilemmas. In *Sociological perspectives on social Psychology*, ed. K Cook, G. A. Fine, & J. S. House, pp. 311-335. Boston: Allyn & Bacon
- Yamagishi, Toshio, & Kikuchi, Masako. 1997. An unpublished experiment on the accuracy of the prediction of the partners' behavior. Faculty of Letters, Hokkaido University, Sapporo, Japan.
- Yamagishi, Toshio, Kikuchi, Masako, Kiyonari, Toko, & Kosugi, Motoko. 1997. An unpublished experiment on the perception of cooperators and defectors. Faculty of Letters, Hokkaido University, Sapporo, Japan.
- Yamagishi, Toshio, & Kiyonari, Toko. 1997. An unpublished experiment on the selection of simultaneous and sequential games. Faculty of Letters, Hokkaido University, Sapporo, Japan.
- Yamagishi, Toshio, & Sato, Kaori. 1986.

Paper presented at the Russell Sage Trust Conference, November 14-16, New York.

Motivational bases of the public goods problem. *Journal of Personality and Social Psychology* 50, 67-73.