SEPARATING TRUST FROM COOPERATION IN A DYNAMIC RELATIONSHIP

PRISONER'S DILEMMA WITH VARIABLE DEPENDENCE

Toshio Yamagishi, Satoshi Kanazawa, Rie Mashima and Shigeru Terai

ABSTRACT

In this article we introduce a new experimental game called Prisoner's Dilemma with Variable Dependence (PD/D), which allows players to separate their trust in their exchange partners from their cooperation with them in an ongoing relationship. The game allows researchers to observe the emergence of trust and cooperation separately, and ascertain the causal relationship between them. In six studies that use the PD/D design, we find that the players of PD/D consistently achieve very high cooperation rates, sometimes mean cooperation rates of about 95%, which are higher than in standard PD games sharing similar design features. These findings demonstrate that separating trust from cooperation is critical for building trust relations. They also show that the GRIT (Graduated Reciprocation In Tension reduction) strategy helps build such relations in the absence of mutual trust. Our results suggest that it is cooperation which leads to trust, not the other way around.

KEY WORDS • trust • cooperation • prisoner's dilemma • risk taking • trust game

Introduction

Interest in trust and cooperation has dramatically increased in psychology, sociology, economics, and related fields in the last few years. Just since 2001, PsycInfo culls 3461 references with the keyword 'trust', and 2169 with 'cooperation' (as of October 2004). A comparable search on EconLit for the same time period produces

Rationality and Society Copyright © 2005 Sage Publications. Vol. 17(3): 275–308.
www.sagepublications.com DOI: 10.1177/1043463105055463

616 references with 'trust' and 1409 with 'cooperation'. Various writers have implicated trust as a necessary ingredient for economic prosperity (Fukuyama 1995; Knack and Keefer 1997), democracy (Putnam 1993; Braithwaite and Levi 1998), and, of course, cooperation itself (Gambetta 1988; Cook 2001). Yet the precise relationship between trust and cooperation, whether trust leads to cooperation or the other way around, remains elusive, and leading theorists disagree on the causal direction (Hardin 2002; Macy 2002).

This disagreement is largely due to the fact that investigators of trust and cooperation often treat the two concepts interchangeably. For instance, the iterated Prisoner's Dilemma (PD) game, the most popular experimental paradigm for studying trust and cooperation, entirely conflates the two. A player in a repeated PD, faced with the binary choice of cooperation or defection, cannot show trust for the other player without cooperating and cannot show distrust without defecting; one cannot cooperate but distrust, or trust but defect.

In this article we introduce a new experimental game called 'Prisoner's Dilemma with Variable Dependence', or PD/D, which allows the players to make separate decisions to trust and to cooperate, and which thus allows us to study the two processes separately. We will present the initial results from a series of experiments with PD/Ds, which demonstrate that subjects can sustain much higher levels of cooperation (often above 90%) than in ordinary iterated PDs, when they can separate their trust from their behavioral choices. In PD/Ds players can initially cooperate without trust, thereby minimizing the negative externalities of potential betrayal by the other player; when cooperation is reciprocated, mutual trust eventually emerges and leads to sustained trustful cooperation. These results therefore allow us to specify the precise mechanism by which *initial cooperation without trust leads to trustful cooperation*. The findings reported later also demonstrate that cautious unconditional cooperators (those who cooperate unconditionally but adjust their levels of trust for the other player) outperform those who employ contingent strategies like TFT. We will discuss the social welfare implications of our findings.

Trust and Cooperation in Experimental Research

There is no consensus on the definitions of trust, and mutually inconsistent technical definitions, for instance trust as a psycho-

logical state or as choice behavior, abound in behavioral sciences (Kramer 1999: 571–4). Worse yet, many scientists use the term for its everyday meaning without clearly defining it. We define *trust*, or to be precise, an act of trust, as an act that voluntarily exposes oneself to greater positive and negative externalities by the actions of the other(s). This is in fact the definition of trust adopted in the trust game literature (Dasgupta 1988; Kreps 1990). Confiding in a friend about a personal problem, loaning money to a colleague (without a legally binding loan agreement and a collateral), and asking an unknown fellow theatergoer to watch a coat on your seat while stepping out to go to the restroom are all instances of trustful behavior. In each case, if the trust placed in the other actor is reciprocated by trustworthy behavior, then the actor is better off than before as a result of positive externalities of the other actor's cooperation. The actor is better off having placed the trust than not having done so. For example, the actor may receive friendly and useful advice from the confidant (or at least have someone to talk to) about the personal problem, generate a reputation as a good colleague and expect the financial favor to be returned in the future, or have the opportunity to go to the restroom without having to carry the coat or losing the seat.

However, trustful behavior necessarily carries the risk of negative consequences if the trust is misplaced and met by untrustworthy behavior. The confidant might break the confidence and share the confidential information with others as gossip, the borrower might default on the loan and not repay it, and the fellow theatergoer might walk away with the coat. In each case, if the trust is misplaced, the actor would have been better off not placing the trust at all. Having to deal with the personal problem alone is better than becoming a laughing stock among friends, being known as a stingy colleague is better than losing a large sum of money, and having to go to the restroom with the coat and losing the seat in the theater is better than losing the coat. The possibility of betrayal (misplaced trust met by defection) makes placing trust in another actor always a risky or even uncertain proposition.

Cooperation is an act that increases the welfare of the other(s) at some opportunity cost where the former is greater than the latter. The forgone opportunity cost (potential gains from defection) is the hallmark of cooperation; without it, an act does not represent genuine cooperation. In the above examples, the confidant forgoes the opportunity to spread juicy gossip about the actor among mutual friends,

the borrower forgoes the opportunity to keep the money by defaulting on the loan, and the fellow theatergoer forgoes the opportunity to possess a new coat. Defection results when the other actor succumbs to the temptation not to forgo the opportunity cost.

The *welfare* of an actor is a multiplicative function of trust and cooperation: welfare = trust × cooperation. The more trust the actor places in the other, the greater the positive externalities of the other's cooperation (and thus the actor's welfare), but at the same time the greater the negative externalities of the other's defection. Further, the actor's trust in the other is a prerequisite for the other's cooperation (or defection). If the actor does not trust, the other actor cannot choose to cooperate or defect. The confidant cannot choose to keep or break the confidence if the actor does not confide, the borrower cannot choose to repay or default on the loan if the actor does not loan the money, and the fellow theatergoer cannot choose to watch the coat or walk away with it if the actor does not ask for the favor.

All of these considerations point to the need to study trust and cooperation separately, to examine the causal effect of one on the other. Given our definitions of trust and cooperation above, however, there is no element of trust in a one-shot (non-iterated) Prisoner's Dilemma (PD) game that is simultaneously played by the two players. This is because a player's choice between C and D cannot affect the size of externalities of their partner's choice.

Trust becomes confounded with cooperation in a sequentially played one-shot PD, in which player A makes a decision between C and D, and then player B, after being informed of the decision of A, makes the same decision. (Both A and B make the decision only once.) This is because the choice for the second player, B, is now between reciprocating A's choice and simple-mindedly pursuing her own self-interest. Which choice, C or D, produces a better outcome for A depends on whether B is a reciprocator or an economically rational actor. If B is a reciprocator, A should choose C since her choice of C makes B also choose C, and A's outcome is then 3 (in a PD in which $T = 3$, $R = 2$, $P = 1$, and $S = 0$). If A chooses D, B will also choose D, and A's outcome is 2. On the other hand, A should choose D if B is an economically rational actor, since B always chooses D regardless of the choice of A. By choosing D, A gets 2, whereas she gets 1 by choosing C. Without knowing exactly which type of a person B is, A has to choose whether or not to trust B to be a reciprocator. The difficulty here

is that A may choose C because A trusts B to be a reciprocator or because A is an altruist who cooperates regardless of the expected choice of B. Similarly, A may choose D because she does not trust B to be a reciprocator or because she is a competitor who is motivated to outperform B, disregarding the absolute level of outcome. There is a large body of experimental evidence that players of one-shot PD games do behave reciprocally (to the expected choice of their partners) and expect their partners to behave reciprocally, even in simultaneously played one-shot games (Hayashi et al. 1999; Kiyonari et al. 2000). This implies that the choice of cooperation and defection in PD games can be partly based on the players' motivation to cooperate, to reciprocate, and their trust in their partners to be reciprocators. In short, both trust and cooperation can potentially be operating behind PD players' choices, and it is difficult to isolate one from the other.

Defection out of fear (the prospect of being exploited by a defecting partner) is a case in point. Player A who cannot trust Player B may still want to cooperate in order to signal her own trustworthiness to B, yet must nonetheless defect in order to protect herself from negative externalities of B's choice and not to receive the 'sucker's payoff' when B defects. Pruitt and Kimmel's (1977) review of over 1000 studies of PD by the mid-1970s concludes that the lack of trust in the partner, rather than greed, constitutes the major factor preventing PD players from taking a cooperative choice. From B's perspective, however, A's behavioral choice of defection is identical to what she would have done if she were driven by greed and the desire to free ride. From A's perspective, her defection is necessitated by B's perceived untrustworthiness, but from B's perspective, A's defection is a proof of *her own* untrustworthiness. Such miscommunication easily leads to mutual recrimination, the 'echo' of mutual defection, in iterated PD games (Axelrod 1984). Further, the experimenter cannot tell from the players' actions whether the mutual defection results from the players' own untrustworthiness (preference for defection) or their inability to trust each other despite their own trustworthiness (preference for cooperation).

Prisoner's Dilemma with Variable Dependence (PD/D)

Game theorists who realize the undesirable implications of conflating trust and cooperation prefer the trust game (Dasgupta 1988;

Kreps 1990) to the PD in studying trust and cooperation separately. The trust game (TG) is played by two players, A and B. Player A is given a choice between trusting (T) and not trusting (NT). If A does not trust B, then the game ends there and the result is a status quo with no changes in the player's welfare. Since the actor's trust is a prerequisite for the other's cooperation, B is given a choice between cooperation and defection only when A chooses to trust B. When A trusts B, then B is given a choice of cooperation and defection. B's choice of defection (not to honor A's trust or NH) produces greater personal welfare for him than his choice of cooperation (to honor A's trust or H): NH > H for B. The confidant forgoes (H) the opportunity to spread juicy gossip (NH), the borrower forgoes (H) the opportunity to keep the money by defaulting on the loan (NH), and the fellow theatergoer forgoes (H) the opportunity to possess a new coat (NH). For A, however, the welfare provided by B's defection is less than that of the status quo, whereas the welfare provided by B's cooperation is greater than status quo. TG thus nicely captures all the elements involved in the earlier examples of trust and cooperation. According to the logic of backward induction used by game theorists, B should not honor A's trust since NH > H, and A, knowing this, should not place trust in B. While successfully capturing the critical elements involved in trust and cooperation, TG has two limitations: it is *static,* and *one-sided.*

The first limitation is recently removed when researchers started using repeated TG rather than one-shot TG. An early example of repeated TG is the Centipede Game introduced by Rosenthal (1981). In the Centipede Game (CG), two players keep choosing between Take and Pass. Each time a player Passes, the total amount of money to be split between the two players increases, and the other player is given another choice between Take and Pass. The game ends when one of the player Takes; then, that player earns the lion's share of the total amount of money and the other player earns very little. Thus, each segment of the CG represents a one-sided TG. One important aspect of CG is that trust and cooperation are inseparably confounded. As McKelvey and Palfrey (1992) point out, Take in CG may be a means to maximize a player's own self-interest or may be lack of the (psychological) trust that the other player will Pass. Similarly, the player may Pass because she is an altruist who wants to improve the other player's welfare (i.e. Pass is based on cooperation using our terminology) or based on her trust that the other player will also Pass and thus she can earn a

better payoff later. CG is thus not appropriate for studying the dynamic relationship between trust and cooperation.

Bolton et al.'s (2003) study, in one condition, development (or, more precisely, maintenance) of trust and trustworthiness by letting their subjects play TG repeatedly between the same two players. They also lift the second limitation of TG when they let the players alternate between the roles of truster and cooperator. Their subjects started with and sustained very high rates of trusting behavior (83%) and cooperation (89%).

Another variant of TG is the investment game or IG (Berg et al. 1995). IG is also played between two players, A and B. As in TG, Player A decides to trust or not to trust B, and B decides to honor A's trust or not. The difference between TG and IG is in the nature of A's and B's choices. In TG, both A and B make binary choices, A between trusting and not trusting, B between honoring and not honoring A's trust. In IG, they make continuous choices: Player A decides how much trust she places in B, and Player B decides how much of the trust placed by A to reciprocate. Specifically, Berg et al. (1995) provided A and B with an endowment of $10 each, and asked A to transfer any of the $10 to B. The money transferred to B is tripled by the experimenter. If, for example, A transfers $4 to B, B receives $12. Player B, who receives the transferred and tripled money in addition to the endowment of $10, then decides whether to send some money back to A. IG thus captures the same elements of trust and cooperation as TG with an additional benefit of allowing the researcher to study varying levels of trust and cooperation. Berg et al. (1995) use IG in a static and one-sided manner, however. It has *not* been used to study development of mutual trust and cooperation between the same two partners. The PD/D we are introducing in this article may be conceived of as a mutual and repeated version of the investment game.

At about the same time as Berg and her colleagues were developing the investment game, and long before economists started applying TG in mutual and repeated manner, we were independently developing a new game, PD/D, that could be used to study development of trust relationship between partners. The game allows each player separately to choose the level of trust she wants to place in the other player, and the behavioral choice to cooperate or defect with the other. In PD/D, it is possible for players to choose to cooperate with the others without trusting them. Further, and more importantly, the game allows researchers to observe the emergence

of, and changes in, mutual trust between players apart from their behavioral choices to cooperate with or defect on each other.

In PD/Ds, we measure A's trust in B as the extent to which A voluntarily chooses her outcome (welfare) to be dependent on B's behavior, or the amount of 'fate control' (Thibaut and Kelley 1959) that A voluntarily gives B. If A trusts B, she voluntarily chooses her outcomes to be dependent on B's behavior and allows him to exercise greater fate control over her outcomes. We will describe the two different types of PD/D with two different operationalizations of dependence, and summarize the initial results from experiments. All experiments reported here are conducted in a fully computerized laboratory. Subjects participate in the experiment in isolation, each in a small room equipped with a computer. Instructions, information, and decision prompts are displayed on the subject's computer monitor, and the subject's decision is entered into the computer, which is connected to the control computer and other subjects' computers through LAN.

Most of these experimental studies have been published in Japanese and thus have not been accessible to anyone who does not read Japanese. The main purpose of this article is to extract major findings that have been discovered by the use of the PD/D experimental design and discuss some of their implications. In addition, it provides non-Japanese-speaking researchers with research findings that have not before been made available in English. We take this opportunity to re-analyze the data when our discussion requires analysis that is not reported in the original studies.

PD/Dm: Trust via Matrix Changes

In the first type of PD/D, subjects choose the amount of their dependence on the other player by choosing a particular payoff matrix, which determines the joint outcomes of their behavior. Subjects can choose to gradually increase or decrease their dependence on the other player by increasing or decreasing the absolute size of their payoffs associated with each outcome, thereby increasing or decreasing their 'fear' of exploitation and temptation to free ride.

In the first round of the iterated PD/D*m* (trust via matrix changes), the two players play the game with the initial symmetrical payoff matrix ($T = 30$, $R = 10$, $P = -10$, $S = -30$; see Figure 1). Then, in the second round, before they make their behavioral

choice to cooperate or defect, each player can choose to increase or decrease their dependence on the other player (or remain at the current level of dependence). If Player A chooses to *increase* her dependence on Player B, for instance, her payoff matrix will shift to the right and the absolute values of all of her payoffs increase by 10% of the original values. Her payoffs in the second round will thus be: $T = 33$, $R = 11$, $P = -11$, $S = -33$. Note that Player A's payoff in the case of mutual cooperation has increased, but the potential damage by Player B's betrayal has also increased. By choosing to increase her dependence on B, A has increased both the positive and negative externalities of B's actions on her. Such a choice requires trust in B that B will cooperate. After choosing the payoff matrix, both players make the binary behavioral choice to cooperate or defect.

If, on the other hand, Player A chooses to *decrease* her dependence on B, then the payoff matrix for the second round will shift to the left and the absolute values of all of her payoffs decrease by 10%.[1] Her payoffs in the second round will be $T = 27$, $R = 9$, $P = -9$, $S = -27$. Player A has chosen voluntarily to decrease her payoff in the event that her partner cooperates (when her trust in B is met with his cooperation), but at the same time decreasing the damage that B can potentially cause her when her trust is misplaced. By choosing to decrease her dependence on B, A has decreased both the positive and negative externalities of B's actions on her. The two players' behavioral choices are informed to them before they decide to increase or decrease their matrix size in the next round.

After each round, Player B can similarly choose to increase or decrease his dependence on A. However, note that his decisions are not known to Player A. Player A only knows of B's behavioral choice (cooperation or defection), but not how much he trusts her. That is why B's payoffs in all matrices except the initial one have a '?' in them. Player A knows that B's payoffs can be more or less than the initial values, but not the current sizes of B's payoffs. We designed our first PD/D*m* this way so that the two players cannot inadvertently communicate their expectations of each other's behavior through their levels of trust; in this version of PD/D*m* (trust via matrix changes), the two players can communicate their trustworthiness only by their own behavioral choices to cooperate or defect. We change this feature of the experiment in PD/D*c* discussed below, however. The design of our PD/D*m* is essentially the same as Van Lange and Visser's (1999) experiment on 'locomotion' (which is

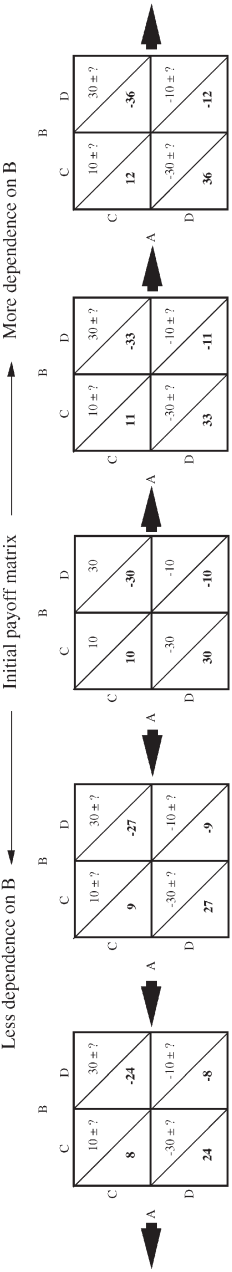

Figure 1. Prisoner's Dilemma with Variable Dependence (PD/D*m*)
(Changes in Player A's payoffs)

their term for the choice of matrix and thus of the level of dependence on the other player). However, our PD/D*m* was invented independently of Van Lange and Visser (1999) and was first published in a Japanese journal a couple of years before theirs (in Kakiuchi and Yamagishi 1997).

In an experiment (Kakiuchi and Yamagishi 1997; first experiment in Yamagishi and Kakiuchi 2000, in English) using PD/D*m*, subjects are first grouped into 'high trusters' and 'low trusters' depending on their responses to two questions on the pre-experimental questionnaire (median-split), administered to potential participants at least a few weeks before the experiment. These questions measure their levels of general trust: 'most people are basically honest,' and 'most people are trustworthy.' Then, two players, unbeknownst to them, are matched on their level of general trust. In the PD/D condition ($n = 40$), the subjects play the PD/D*m* described earlier, and in the PD condition ($n = 40$), they play the ordinary PD with the initial payoff matrix on all rounds without the ability to increase or decrease their dependence by the choice of payoff matrix. Subjects in both conditions play the game for 48 rounds (although they do not know the exact number of rounds ahead of time). The experiment takes place in Japan with Japanese subjects. Subjects interact with their partners in their own compartment, and they have no opportunity to meet other participants either before, during or after the experiment. Their decisions are completely anonymous. They start with the matrix located at the center of Figure 1, and play the game for 48 trials. They are paid one yen for each point they earn in the experiment.

In the PD/D condition, both high trusters and low trusters increase their levels of dependence on the other player throughout the experiment, although high trusters do so much more quickly than do low trusters. Kakiuchi and Yamagishi (1997) break down the 48 trials into six trial blocks of eight trials. The difference in the matrix size between low trusters and high trusters started to emerge in the fourth trial block in which the increase in the matrix size was 145.3% of the original matrix among high trusters and 96.8% among low trusters. During the last trial block, the increase reached 254.5% of the original matrix among high trusters, and 133.4% among low trusters. As Table 1 shows, the mean level of cooperation in the PD/D condition is higher among high trusters than among low trusters, whereas the difference between low and high trusters is small in the PD condition. More importantly, however, regardless

Table 1. Trust and cooperation in Prisoner's Dilemma (PD) and Prisoner's Dilemma with Variable Dependence (PD/D) games

Experiment and subjects	*Condition*	*n*	*N of trials*	*Trust*	*Cooperation***
KY97	PD high trust	20	48		.45
Japanese	PD low trust	20	48		.53
	PD/D*m* high trust	20	48	24.42*	**.82**
	PD/D*m* low trust	20	48	18.80*	**.49**
MY2001	Random PD	150	24		.40
Japanese	PD	52	36		.66
	PD/D*c*	98	36	7.35 coins	**.76**
CYCCMM2003	Random PD	106	25		.39
Americans	PD	32	45		.58
	PD/D*c*	28	45	8.92 coins	**.90**
	Random PD/D*c*	46	45	6.81 coins	**.53**
Japanese	Random PD	192	24 or 25		.42
	Random PD/D*c*	42	45	5.06 coins	**.47**

TMY2003a Japanese	PD/D*cc*	40	22 minutes	.99	**.95**
TMY2003b, Experiment 1 Japanese	*n*PD/D*cc* Naïve subjects	14	45 minutes	.93	**.83**
	*n*PD/D*cc*, Naïve and expert subjects	11	45 minutes	.99	**.79**
TY unpublished Japanese	*n*PD/D*cc* Cost Condition	14	18 minutes	.97	**.91**
	*n*PD/D*cc* No-cost Condition	14	18 minutes	.98	**.91**
MYM2003 Japanese	*n*PD/D*cc*	38	30 minutes	.97	**.89**
Americans		44	30 minutes	.98	**.84**

* Average net increase in matrix size.
** Proportion of the times entrusted money is returned for PD/D*c* and PD/D*cc*.

of their level of general trust, players of PD/D are able to sustain much higher levels of mutual cooperation than players of PD, $F(1, 36) = 4.21$, $p < .05$. The mean cooperation rate among high trusters in PD/D is 82%, whereas that among low trusters is 49%, about the same as the mean cooperation rates of high and low trusters in the PD condition (45% and 53%). The average cooperation rate in the last trial block reaches 88% among high trusters and 67% among low trusters in PD/D, whereas the average cooperation rate in PD in the last trial block is 72% among high trusters and 73% among low trusters. It appears that the ability to separate players' behavioral choices from the amount of trust they place in the other player increases the likelihood of mutual cooperation, especially when high trusters play each other (the interaction effect of general trust and game type, $F(1, 36) = 6.33$, $p < .05$). Van Lange and Visser (1999), who used a similar design, also report a high level of cooperation and reach the same conclusion.

PD/Dc: Trust via Coin Entrustment

Upon completion of the first experiment with PD/D*m*, we reached the conclusion that the experimental task for the subjects, where they make separate decisions first to choose the payoff matrix (only one half of which was visible to them) and then to make the behavioral choice to cooperate or defect, might have been cognitively too demanding for our subjects. We have therefore designed an entirely different experiment, which nonetheless derives from the same underlying logic and thus represents a conceptual replication of our first experiment.

In the PD/D*c* (trust via coin entrustment), each player first receives an endowment of 10 coins at the beginning of each round, then each makes a decision as to how many of the 10 coins they want to entrust to the other as in the investment game. However, PD/D*c* differs from IG in that it is symmetrical; each player makes the trusting decision. They can choose any number of coins, shown on their computer display, between 0 and 10 to entrust (the minimum number of coins to entrust is one in some experiments) and keep whatever coins they do not entrust to the other. The number of coins entrusted by the subject appears on her partner's computer display. Then, in the second phase of the round, the players make the binary decision as to whether or not to return the entrusted coins to the other player. If they choose to return them (equivalent to

cooperation), the coins are doubled in number before being returned to the original player. If they choose not to return them (defection), they get to keep the entrusted coins, but they are not doubled in number. Earlier experiments on PD (Yamagishi and Sato 1986; Kollock 1993) have allowed subjects to choose continuous levels of cooperation (rather than the binary cooperate–defect decision). However, none of these experiments allow their subjects to separate trust from cooperation.

PD/D*c* is similar to IG in some respects but different in others. The two share the same idea that the trusting choice creates an opportunity for the trusted to cooperate or defect. The amount of trust determines the size of the pie to be shared and cooperation divides the pie. An important difference exists, however, between the two games that affect interpretation of the subject's behavior. In IG, the money transferred from Player A triples in value *before* it reaches Player B; in PD/D*c*, the coins entrusted to Player B by Player A double in value *only when* Player B chooses to return them. In IG, it is the truster's decision that increases the size of the pie to be divided; in PD/D*c*, it is the trusted player's decision. From the point of view of the trusted, how much to return to the truster in IG involves a division of a fixed sum; it may thus reflect his fairness concerns, but it is difficult to call it cooperation. In contrast, the choice for the trusted in PD/D*c* is between greater collective welfare versus maximization of own individual welfare disregarding the collective welfare. IG is thus better suited to study fairness behavior on the part of the trusted and expectation of fairness on the part of the truster. PD/D*c* is better suited to study cooperation on the part of the trusted and expectation of cooperation on the part of the truster. We use PD/D*c* since we are interested in the relationship between trust and cooperation rather than between trust and fairness.

The second important difference between the two games is whether the choice of the trusted is binary or continuous. (The choice of the truster is continuous in both games.) When the game is asymmetrical, as in the original IG, the trusted's choice must be continuous to allow him to behave in a fair manner. For example, let us assume that the truster sends all \$30 to the trusted, and that the trusted has a binary choice of returning all \$30 or keeping all \$30. Then, the choice for the trusted is between two kinds of unfairness. When he sends back the money, the truster gets \$30 and the trusted \$10, whereas when he does not send back the money, the

truster gets nothing and the trusted gets \$40. This problem disappears, however, when the game is symmetrical; if both transfer all of their endowments to their partners and each chooses to send back all, each gets \$30. We use PD/D*c* for its simplicity.

In PD/D*c*, the choice in the entrustment stage is continuous (the players can choose to entrust any number of coins), whereas the choice in the behavioral stage is binary (they can either return or keep the entrusted coins *in toto* but may not return some of the coins and keep others). The game is symmetrical; each player first makes a choice of how many coins to entrust to her partner, and, second, whether or not to return the entrusted coins. The player's total earnings for each round is the sum of the number of coins she chooses not to entrust to the other player in the entrustment stage, plus the number of coins entrusted to her if she chooses not to return them, and double the number of coins that she entrusts to the other and are returned to her (if they are returned by the other player) in the behavioral stage. The payoff structure of PD/D*c*, once the decisions of how many coins to entrust to the other player has been made, is thus: $T = 10 + m + n$, $R = 10 + n$, $P = 10 + m - n$, $S = 10 - n$, where n is the number of coins a player entrusts to the other player and m the number of coins the other player entrusts to him.[2]

Comparison of PD with PD/Dc. The first study using the PD/Dc is conducted by Matsuda and Yamagishi (2001), in which cooperation rates are compared between the PD/D condition and the PD condition. Ninety-eight Japanese subjects are assigned to the PD/D condition and 52 to the PD condition. In order to make the results comparable, PD players also choose between keeping and returning the coins entrusted by the other player. However, unlike in the PD/D*c*, players in the PD condition cannot choose how many coins to entrust; the number of coins to be entrusted to the other player is randomly determined by a computer program and there is nothing the players can do to alter their 'trust' level. The players in the PD condition thus only make the behavioral decision to return (cooperate) or keep (defect) the coins. All subjects first experience the random-partner PD condition for 24 rounds. In this initial random-partner PD condition, pairs are randomly formed each time among six or eight subjects, and they are not informed of the identity of their partner. The first 24 rounds thus represent repeated one-shot PD games. The number of coins to be entrusted to the

other player is randomly determined by the computer between 1 and 10 (uniform distribution). Then, at the beginning of the 25th round, they are matched with one partner based on their cooperation rates in the first 24 rounds, and they play for the remaining 36 rounds with the same partner. The participants are not informed of the total number of rounds in advance. The manipulation of the two games takes place during the latter 36 rounds. Subjects in the PD/D condition receive additional instructions before the 25th round and are told that they are now matched with one partner for the remaining rounds, and that they can choose for themselves how many coins to entrust to their partner. Subjects in the PD condition receive instructions only about permanent pair matching. Each coin they earned in the experiment is converted to 2 yen at the end of the experiment and is paid to the subject as the reward for participation.

Figure 2 presents the results of this experiment. The mean cooperation rate during the first 24 rounds (i.e., in the random-PD

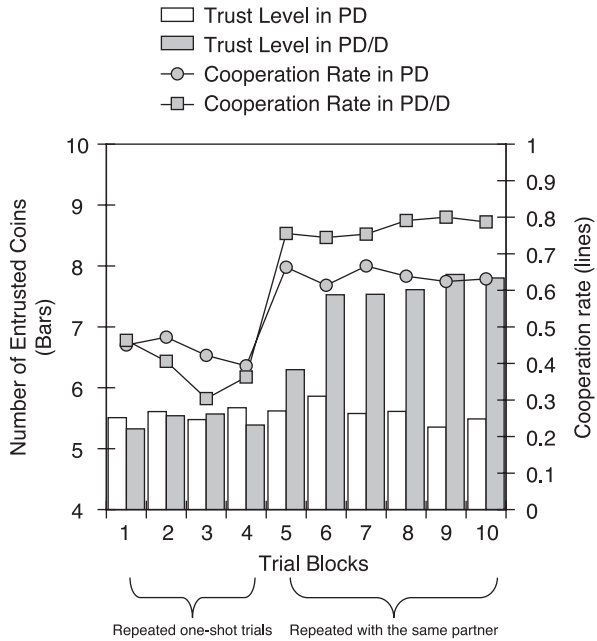

Figure 2. Trends over trial blocks of trust (number of entrusted coins) and cooperation (proportion of coins returned by the trusted player) in the PD and the PD/D condition.

condition) is 40%. When two subjects are permanently matched in the PD condition (during the last 36 rounds), this cooperation rate improves to 66%. This improvement in cooperation rate reflects a well-established difference between one-shot games and repeated games (Axelrod 1984); the cooperation rate in PD games is higher when the game is repeatedly played between the same two people than when it is played once. More importantly, the mean cooperation rate among the PD/D players of 76% is even higher than that observed among the PD players under similar conditions, $F(1, 148) = 3.16$, $p < .08$.[3] The results of this PD/D*c* experiment replicate the earlier results with PD/D*m* that allowing subjects to separate trust and cooperation decisions significantly increases the level of cooperation among them.

Comparison of Japanese and American subjects. Since the two studies described earlier use Japanese students as subjects, Cook et al. (2005) run a replication experiment in the USA and Japan using American and Japanese[4] subjects to replicate the finding that separating trust from cooperation facilitates cooperation in continuing relations. In addition, they examine the hypothesis that the cooperation-enhancing effect of separation of trust from cooperation will be more pronounced among Americans than Japanese.

This hypothesis is derived from the consistent finding that the Japanese are more risk averse than Americans. For example, according to Hofstede (1980), Japan is near the top of the 40 nationalities he studied on their level of uncertainty avoidance whereas Americans are near the bottom of the list. Yamagishi et al. (1998) report findings from two cross-societal experiments in which they found that the Japanese, compared with Americans, are less willing to engage in cooperative behavior in the absence of a collective system that reduces social uncertainty, and are more willing to engage in behavior that reduces social uncertainty. The results of experiments provide strong evidence for the argument that makes a distinction between commitment formation as uncertainty avoidance and trust as risk taking. High trusters – those who were shown to have a high level of general trust in their responses on a trust scale – are less likely to engage in commitment formation than are low trusters when faced with a socially uncertain situation (Yamagishi et al. 1998). In uncertain situation, those who prefer to form a commitment relation with a particular partner and thus reduce the risks within such a relation are less trustful of other people in general.

The same result concerning the effect of general trust on commitment formation was obtained both with American and Japanese participants. Those low in general trust of others in both societies are more likely to form commitment relations with those who are trustworthy despite the opportunity cost involved. The predicted difference between the Japanese and the Americans reflects not only relative levels of uncertainty avoidance, but also different levels of general trust (typically higher in the USA than in Japan).

The same computer program (with translated display materials) that has been used by Matsuda and Yamagishi (2001) is used in this replication study with 60 American subjects, and the results are compared to the Japanese results. In the random-PD condition (first 24 or 25 rounds), the cooperation rate is higher among Japanese subjects (40%) than among American subjects (30%), and the difference is statistically significant, $F(1, 206) = 8.75, p < .01$ (see Table 1). The same difference is also observed in the PD condition (last 35 or 36 rounds), where the average cooperation rate is 66% among Japanese and 58% among American subjects, though the nationality difference is not statistically significant, $F(1, 82) = 1.17$, *ns*. This difference reverses itself in the PD/D condition in which subjects can choose the level of trust independent of the choice between cooperation and defection. American subjects in the PD/D condition entrust an average of 8.92 coins per round, which is significantly greater than the average number of coins Japanese participants entrust to their partners (7.35), $F(1, 124) = 7.98, p < .01$. The mean cooperation rate among American subjects jumps from 58% in the PD condition to 90% in the PD/D condition, while the same increase among Japanese subjects is from 66% to 76%. Given the separation of trust from cooperation, Americans cooperate at a higher rate than Japanese, and the nationality difference is statistically significant, $F(1, 124) = 5.67$, $p < .05$. As predicted by Cook and her colleagues, Americans are better able than Japanese to take advantage of the opportunity to separate trust and cooperation provided by the PD/D; the interaction effect on cooperation between nationality and game type is significant, $F(1, 206) = 5.67$, $p < .05$. The main effect of the game condition (PD versus PD/D) in the nationality $\times$ game condition ANOVA of the cooperation rate during the last 36 rounds is highly significant, $F(1, 206) = 19.76$, $p < .0001$, whereas the main effect of nationality is not, $F(1, 206) = 0.42$, *ns*.

Cook and her colleagues examined another condition of PD/D in which two players were randomly matched in each trial, and found that the positive effect of the opportunity to separate trust and cooperation provided by the PD/D requires that the game be played between particular partners repeatedly. While American subjects entrusted more (6.81) coins to their partners than did their Japanese counterparts (5.06 coins), and the difference was statistically significant, $F(1, 86) = 11.58$, $p < .01$, their cooperation level was not as high as in the PD/D with a fixed partner (see Table 1). These results indicate that the positive effect of separating trust and cooperation exists only when it is possible to establish a trust relationship between particular partners in which the two players trust each other and cooperate with each other.

PD/Dcc in a Continuous Decision Environment

In the continuous-decisions version of the PD/D*c*, we eliminate discrete decision rounds. Players receive an endowment (real money rather than coins) at some unpredictable and irregular interval (e.g. every two to three minutes) and the game continues uninterrupted and unpunctuated for some period of time unknown to them (e.g. 45 minutes). Within this time period, players can freely choose to entrust, return or keep the endowment in discrete units at their leisure. There are no time limits; once money is entrusted to them, for example, they can indicate their choice to keep or return it immediately, or they can postpone the decision for as long as they want.

There are three constraints on players' behavior in the continuous-decision version of PD/D*c*, which we will call PD/D*cc*. First, once they entrust money to a player, they cannot further entrust more to the same player before she decides whether to return or keep the money entrusted by them earlier. Second, if Player A entrusts money to Player B, Player B must first decide whether to keep or return it before entrusting his own money to Player A. Two players cannot simultaneously entrust money to each other. Third, even though unentrusted money accumulates in the players' accounts without limit, they can entrust only up to the size of each deposit in a single entrustment decision.

In an experiment with PD/D*cc*, Terai et al. (2003a) find that both trust (measured by the proportion of the endowed money entrusted to the partner) and cooperation (measured by the proportion of the

time that players return the entrusted money to the entruster) are *extremely* high (see Table 1). Forty Japanese subjects are paired with each other throughout the entire experiment that lasts for 22 minutes. Subjects, however, do not know how long the experiment lasts. They are given 30 yen from the experimenter every 75 to 105 seconds, and choose to entrust any of that money, in increment of 5 yen, to their partner. They may repeat the entrusting decision as many times as the endowment money is available. They may entrust the whole 30 yen to their partner at one time, or entrust 5 yen, wait until it is returned, and then entrust another 5 yen, wait until it is returned, and so on. Anonymity of their choices is tightly maintained. The subjects in this experiment entrust 99% of the money they are given by the experimenter, and return the entrusted money 95% of the time.

nPD/Dcc in a Continuous Decision Environment with Multiple Partners

Terai et al. (2003b) use the PD/D*cc* in an environment in which players can simultaneously interact with multiple partners. In this *n*-person PD/D*cc* or *n*PD/D*cc*, there are more than two players simultaneously in the game, but, unlike social dilemma games, all transactions are bilateral (only two of the players involved in the exchange of entrustment and return of money at a time). There is no pooling or dividing of resources by the *n* players. However, players can choose to spread their endowments by entrusting them to multiple players *simultaneously* (e.g. entrust 10 yen to Player B, 5 yen to Player C, and 15 yen to Player D).

A total of 14 Japanese students participated in the experiment in two seven-person groups. Each subject receives an endowment of 50 yen every two to four minutes, and decides how much of it (in increments of 5 yen) to entrust and to whom. The experimental session lasts for 45 minutes, though subjects do not know how long it lasts. The subjects in this experiment again show extremely high levels of trust and cooperative behavior (see Table 1). They entrust 96% of the money they are given by the experimenter, and return 93% of the time they are entrusted money by their partner.

Each player has six other players to whom to entrust, but they tend to concentrate on a few. Of the entrusted amount, 32% goes to the top trustee, and 23% goes to the second, together accounting for more than a half of the total entrustment. And the cooperation

rate is especially high among the top trustees. The cooperation rate of the most preferred trustee (not necessarily the same person for all players) is 100%, and that of the second most preferred trustee is above 99%. In short, subjects of this experiment form trust relationships in which near perfect cooperation is ensured.

These results are replicated by a follow-up experiment in which each seven-person group consists of four naive subjects and three experts (graduate students who are familiar with experimental gaming). The expert subjects are paid in the same way as the naive subjects. They are encouraged to do their best to 'exploit' others to make the most money they can. While the use of 'expert' subjects and encouragement for them to take advantage of their expertise reduce both trusting and cooperative behavior, they are still at relatively high levels. The 14 subjects entrust an average of 96% of their endowment, and return 81% of the entrusted money. Interestingly, the average entrustment level (99%) is higher among expert subjects than among naive subjects (93%).

Mixing American and Japanese Participants

Both American ($n = 44$) and Japanese subjects ($n = 38$) participated in an experiment conducted by Mashima et al. (2004), using the same *n*PD/D*cc* as in Terai et al. (2003b). The subjects participate in seven- or eight-person groups, consisting of three or four Japanese subjects and four American subjects. They participate in the experiment in similar isolation rooms in each location, Hokkaido University in Japan and Cornell University in the USA. These two laboratories were connected via the Internet, and the program was controlled from a server computer located in the Hokkaido laboratory. In one condition of the experiment, the subjects do not know that they consist of two nationalities; they have no idea that some of the fellow participants are in another country. In the other condition, subjects know that participants are from two nationalities, and each of the other members displayed on the subject's computer screen comes with a national flag indicating whether the member is Japanese or American. To accommodate 11 hours of time difference, the Japanese participated in the morning and the Americans in the evening.

Mashima and her colleagues (2004) find no effect of nationality information or the nationality of the subject on the trusting

behavior; both American subjects (98%) and Japanese subjects (97%) entrust most of the endowment to the other members of their group (see Table 1). Nor do they find any ingroup-favoring tendency in the choice of trustees. The cooperation rate in this experiment, while still very high, is slightly lower than in the other study (Terai et al. 2003b) with the same design and computer program. The cooperation rate is 89% among Japanese subjects and 84% among American subjects. This difference between the two groups of subjects is statistically marginal, $F(1, 78) = 3.37$, $p < .08$, in a subject's nationality × partner's nationality × display of nationality information ANOVA. The display of nationality information helps improve the cooperation rate, even though the level of cooperation with ingroup members is not higher than that with outgroup members. The cooperation rate is 83% without nationality information and 92% with nationality information, and the difference is significant, $F(1, 78) = 7.67$, $p < .01$.

Introducing Transaction Costs

The extremely high level of cooperation in the continuously played PD/D (PD/D*cc* and *n*PD/D*cc*), even compared to the already high level of cooperation in PD/D*c*, is very impressive. The average cooperation rates in these games are 79–95% among Japanese subjects and 84% among American subjects. The mean trust levels are 93–99% among Japanese subjects, and 98% among American subjects. These extremely high levels of trust and cooperation can be attributed to two factors, in addition to the features of the PD/D games discussed above. The first is the reduction in risk. In the discrete PD/D*c* game, trusting the entire endowment requires a large risk – entrusting 50 yen involves a risk of losing 50 yen. And yet a player has to entrust the maximum amount to maximize the potential gain. If a player entrusts only 5 yen, she can gain only 5 yen when the entrusted money is returned. (Five yen of the returned 10 yen is the player's original money, so she only gains 5 yen from this transaction, not 10 yen). If she entrusts 50 yen, then she can potentially gain 50 yen. In the continuous PD/D*cc*, she does not need to take a large risk for achieving the maximum gain of 50 yen. She may simply entrust 5 yen 10 times, each following the return from the partner.

The data from the experiment demonstrate that subjects are in fact using this strategy of dispersing risk. Subjects in the PD/D*cc* game can minimize risk at no cost. This leads to the second factor – reduction in temptation. Insofar as a player disperses the risk in this way, each time entrusting the minimum possible amount, waiting until the partner returns the money, and then entrusting the minimum amount again and again, the gain for the partner from defection is minimized as well. The most the partner earns from defecting (i.e., not returning) is only 5 yen. The minimum amount of 5 yen is too little for them to risk jeopardizing the relationship and losing a partner who will otherwise keep returning the money he entrusts in the future. The continuous version of PD/D*cc* represents a social situation in which people do not need to trust their partners. Yamagishi and Yamagishi (1994; see also Yamagishi 1998; Yamagishi et al. 1998) call such a social situation one characterized by *assurance* relations rather than *trust* relations.

Terai and Yamagishi conducted an unpublished experiment in which they examined this account of extremely high levels of cooperation in the continuous version of PD/D or PD/D*cc*. In one condition of their experiment, Terai and Yamagishi introduce a transaction cost for each entrusting action. Specifically, the subjects in this condition are charged 5 yen each time she entrusts money to another player. Thus, entrusting the minimum amount of 5 yen is a sure way of losing money. Even when the entrusted person returns the 5 yen and the truster gains 5 yen, 5 yen 'entry fee' is charged by the experimenter and her profit from this successful transaction is zero. Thus, the best she can expect from a trustworthy trustee is zero profit, and the worst is a loss of 10 yen (5 yen of 'waste' and 5 yen of the entry fee). With the entry fee, players have to entrust a larger amount at a time to make the entrusting behavior worthwhile, compromising the risk dispersion strategy. The introduction of the entry fee is thus expected to reduce the levels of trust and cooperation, by making risk dispersion less profitable and thus increasing temptation to keep the larger amount of entrusted money.

Terai and Yamagishi manipulate the cost condition (no cost versus 5 yen for each entrustment) as a within-subjects factor. Specifically, their subjects ($n = 35$) play the same game in the cost condition for the first 18 minutes, and in the no-cost condition for the last 18 minutes. On average, they entrust 98% of the endowed money in the no-cost condition, compared to 97% in the cost condition, and the difference is not statistically significant, $t(34) =$

1.78, *ns*. However, they entrust their money less frequently in the cost condition (21.3 times) than in the no-cost condition (28.3 times), as expected, and the difference is highly significant, $t(34) = 3.86$, $p < .001$. At the same time, they entrust a larger amount per entrustment decision in the cost condition (15.6 yen per entrustment decision) than in the no-cost condition (14.0 yen). The proportion of the times the entrusted money is returned is not different between the two cost conditions. In either condition, the entrusted money is returned 91% of the time. On the other hand, the proportion of the entrusted money returned in the cost condition (87%) is significantly lower than in the no-cost condition (93%), $t(34) = 2.16$, $p < .05$, suggesting that the trusted players kept the entrusted money more frequently when the amount was larger than when the amount was smaller. While the cooperation rate in the cost condition is still very high, it is lower than the cooperation rate in the no-cost condition in which players can disperse risks without cost.

Cautious Cooperation

The difference in cooperation rate between the ordinary repeated PD games and the PD/D games is impressive. Except among the low-trusting pairs, the subjects in the experiments presented above cooperate at a much higher rate in PD/D games than in equivalent PD games. In Kakiuchi and Yamagishi (1997), cooperation rate among high-trusting Japanese subjects increases from 45% to 82%. In Matsuda and Yamagishi (2001), the cooperation rate among Japanese subjects increases from 66% to 76%. The positive effect of separating trust from cooperation is even more pronounced among American subjects. In Cook et al.'s (2005) study, which is a replication of Matsuda and Yamagishi (2001) with American subjects, cooperation rate among American subjects increases from 58% to 90%! The high rates of cooperation in the PD/D conditions in these studies, 82%, 76%, and 90%, compared with those in the ordinary repeated PD conditions, 45%, 66%, and 58%, are impressive. And the cooperation rates are even higher in the continuous version of the PD/D, 83–95% among Japanese subjects, and 84% among American subjects.

Here, we address the question of why separating trust from cooperation is so important in generating and sustaining such high levels of cooperation. We have already provided the clue for this

answer in the introduction. Separating trust from cooperation provides a remedy for 'fearful defection'. In ordinary PD games in which trust and cooperation are conflated, even those who are willing to forgo the opportunity to exploit the partner unilaterally in order to achieve and maintain mutual cooperation may still fail to behave cooperatively to avoid the 'sucker's payoff'. Using Arneson's (1982) terminology, they are 'nervous' or 'reluctant' cooperators. Nervous or reluctant cooperators recognize the importance of mutual cooperation, and are willing to cooperate if they are convinced that others are also willing to do so. In the absence of such assurance, however, they hesitate to cooperate.

There are two possible reasons for this. First, they may be afraid that their efforts for acting cooperatively may be wasted if others do not cooperate. For example, people are reluctant to join the election campaign of a candidate who seems unlikely to attract much support. Arneson (1982) calls these people 'nervous cooperators'. Second, they may be afraid that their cooperative initiatives will be taken advantage of by predatory non-cooperators. For example, a price-fixing cartel (although it is illegal in many countries) will fail if members are afraid that others would undersell. If some corporations lower their prices, those that do not will lose their share and will eventually be driven out of the market. Arneson (1982) calls those who are afraid that they may be exploited by non-cooperators 'reluctant cooperators'. Nervous or reluctant cooperators will cooperate if they are convinced that others will as well. If they cannot trust others, they will not cooperate.

This is the story of PD, but not of PD/D. The nervous or reluctant cooperators do not need to defect out of fear (of waste or of exploitation) in the PD/D. In order to reduce their fear and feel safe, they only need to reduce their levels of trust. In the PD/Dc, for example, nervous or reluctant cooperators can stop sending money to their partner while returning all the money they are entrusted by the partner. This conveys their intentions to build a mutually cooperative relation with the partner, and thus prevent a vicious cycle of self-fulfilling suspicion (or using Lawler's [1986] terminology, a conflict spiral).

We believe that this is an important part of the story, but it is not the whole story. The other part of the story is that a relationship in which both parties can safely trust each other is self-sustaining once it is achieved. It is difficult to achieve such a relation, and thus it is a valuable resource for both parties. A friendship in which

you can safely confide in your friend is hard to develop, and thus such a friendship is extremely valuable to you. Furthermore, such a relationship is more valuable to you to the degree it is difficult to achieve. A mutually trustful and trustworthy relationship is self-sustaining since it is too valuable to destroy for a short-term gain. Conversely, it takes a larger short-term gain to forgo such a valuable relation. The difficulty, of course, is how to build one. Separating trust from cooperation provides a first step toward achieving such a difficult state, but it is just a start. Players of a PD/D game can reduce their fear by not trusting their partners. The initial strategy of a nervous or reluctant cooperator in a PD/D game is cooperation without trust. The challenge is how to transform this mutually distrustful relation into a mutually trustful one.

Matsuda and Yamagishi (2001) analyze the players' strategies, and come to the conclusion that *cautious unconditional cooperation* is the best way to achieve and sustain trustful cooperation. They analyze what kind of strategy each of their subjects uses in their experiment, by classifying them into types based on their trusting and cooperation behaviors in response to their partner's behaviors in the previous round. Then, they compare the earnings associated with each strategy. According to their analysis, the most frequently observed players are the ones who unconditionally cooperate (regardless of the partner's behavior in the previous round) while adjusting their level of trust according to their partner's cooperation–defection choice in the previous round. They start with entrusting only a few coins, and increase the number of coins to entrust when their partner returns them, and decrease it when the partner does not return them. They call these type of subjects 'cautious unconditional cooperators'. They are the highest earning players as well. They ($n = 25$, 30% of all the subjects in the PD/D condition) earn 650.8 yen on average and outperform another major type, double TFTers ($n = 11$, 13%), who adjust their cooperation–defection choices as well as their level of trust to their partner's behavior on the previous round and earn an average of 477.0 yen.

Figures 2 and 3 dynamically represent the strategy of cautious unconditional cooperation in a PD/D*c* across all subjects (Matsuda and Yamagishi 2001). They show that as soon as the subjects are paired in repeated games (at the beginning of Trial Block 5), their cooperation rates immediately soar to a very high level (around 80%); in other words, their cooperation immediately becomes almost unconditional. However, their trust, measured by the number

of entrusted coins, does not accompany the rapid increase in cooperation. In the first trial block after pairing (Trial Block 5), the trust level is still almost as low as in the previous trial blocks during one-shot games; even while their cooperation is nearly unconditional, the players are cautious. Their trust goes up only after a full trial block of six trials, during which their cautious trust was reciprocated by the other player's cooperation. Only then does their trust go up as high as their cooperation. Figure 3 shows the lagged nature in the increase in trust during the fifth and the sixth trial blocks. Further analysis indicates that the gradual increase in trust during the 5th and the 6th trial blocks shown in Figure 3 occurs only among unconditional cooperators who never defect in these two blocks ($n = 45$). Their trust levels increase from 5.36, 6.76, and 7.62 in trials 25, 26, and 27, respectively, to 9.73, 9.78, and 9.96 in trials 34, 35 and36, respectively. Among those who defect at least once in the two trial blocks ($n = 53$), practically no increase in trust occurs; their trust levels in trials 25, 26 and 27 are 4.06, 4.94, and 5.09, respectively, and in trials 34, 35, and 36, they are 5.62,

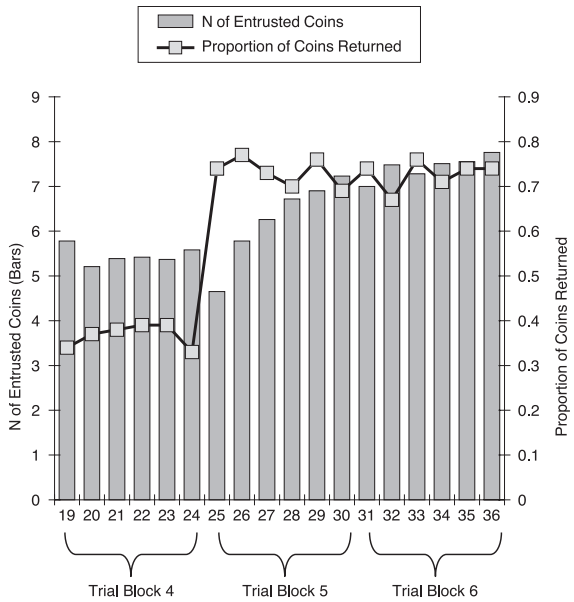

Figure 3. Trends over trials in the fourth through sixth trial blocks of trust (number of entrusted coins) and cooperation (proportion of coins returned by the trusted player) in the PD/D condition.

5.66, and 5.89, respectively. These results clearly demonstrate that cautious unconditional cooperation precedes (and, if reciprocated, eventually leads to) trustful cooperation. *It is cooperation that leads to trust, not the other way around.*

The success of cautious unconditional cooperators demonstrates the importance of cooperation without trust as the first stage of eventual success in formation of trustful cooperation, through the process known as GRIT (Graduated Reciprocation In Tension reduction; Osgood 1962). Lindskold's (1978) review of empirical studies suggests that following the prescriptions set out in the 10 principles of the GRIT strategy leads to building trust and then to cooperative relationships between exchange partners of various kinds. Cautious unconditional cooperation in PD/Ds can accomplish two goals that PDs do not allow. First, unconditional cooperation allows the players to signal their trustworthiness to their exchange partner. Second, unlike cooperation in PDs, cooperation without trust in PD/Ds allows the players to do so without much risk. The only thing that cautious unconditional cooperators lose by cooperating is the opportunity cost of exploiting their exchange partner by defecting on them; they do not risk much of their own resources because the PD/D allows the players to cooperate without voluntarily transferring control of their welfare to their partner. Once their low risk trust (e.g. entrustment of one coin) is reciprocated, then players can gradually increase their level of trust in the other player, by entrusting more and more coins, while at the same time unconditionally cooperating by returning all the coins entrusted to them. Mutual trust, and fully trustful cooperation, will eventually develop out of initial cooperation without trust. Lawler and Yoon's (1993, 1996) work on relational cohesion demonstrates that individuals often experience euphoria ('emotional buzz') when they complete successful exchange via mutual cooperation. Such emotional buzz after successful exchange might be the proximate mechanism that leads to greater trust to the extent that individuals are more likely to trust those with whom they are happy.

General Discussion

Individual welfare, as we noted earlier, is a multiplicative function of trust and cooperation: welfare = trust × cooperation. The level of trust multiplies the beneficial effect of cooperation; the more one

trusts the other, the more beneficial the other's cooperation. The same is true at the macrosocial level; social welfare in society is a multiplicative function of the average level of trust among people and the mean rate of cooperation. The society as a whole benefits when individuals engage in *trustful cooperation*; trust in and of itself does not produce social welfare if unaccompanied by cooperation, and cooperation in and of itself does not produce social welfare if unaccompanied by trust. In PD/D*c*, for example, 100% trust does not increase the number of coins in the society (i.e. the group of subjects) unless they are returned, and 100% cooperation does not increase the number of coins in the society either if nobody entrusts any; in fact, no one even gets a chance to cooperate if nobody entrusts.

There is a paradox here. While maximal social welfare is possible only with mutually trustful cooperation, trust emerges from initial cooperation without trust. Actors must first signal their trustworthiness by unconditional cooperation, but must also begin slowly by gradually and cautiously increasing their level of trust in the others. When and only when their initial low trust is rewarded with cooperation can they begin to trust more. Mutual trust emerges through a series of risk-taking, by the act of trusting more today than yesterday. Once mutually trustful cooperative relation emerges, however, it is self-sustaining because it is difficult to replicate it in other relationships. Given uncertainty inherent in any new exchange relationship, it would be utterly irrational for self-interested actors to defect in a self-sustaining cooperative relationship.

The findings from the series of experiments reported here, all using some forms of PD/D methodology, shed light on the intricate relationship between trust and cooperation. Do we cooperate when we are trusted more than when we are not trusted? Or, alternatively, do we trust when our partner has behaved in a cooperative manner? The success of the GRIT strategy adopted by the cautious unconditional cooperators – those who unconditionally cooperate, but adjust their levels of trust to their partner's level of cooperation – implies the latter causality. It is important to be unconditionally cooperative for the success of the GRIT strategy (Osgood 1962; Lindskold 1978). Unconditional cooperation by a player invites the partner to trust them. Once the process starts, the relationship between trust and cooperation takes a continuous spiral. As the partner increases the level of his trust, the relation becomes more valuable to him and, as a result, he is less tempted to defect. The GRIT strategy and the

unconditional cooperation is needed, however, to increase the level of trust gradually to the self-sustaining level through the spiral of mutual trust and cooperation.

At least in the initial stage, cooperation gives rise to trust, not the other way around. Separating trust from cooperation is especially useful in this early stage of trust building, since it allows players to be consistently cooperative while minimizing the risk of exploitation by their partners. The double-TFT strategy that reciprocates in kind – defection and reduced trust for defection, and cooperation and increased trust for cooperation – is not a very useful strategy in this early stage, since it destroys the partner's fragile trust, which is needed to get the whole process going. Turning the other cheek can be the best strategy when your partner has no trust in you, provided that you protect your cheek well so that your enemy's slap on your cheek won't hurt you much.

We end with two implications of our research with PD/D. First, our findings suggest the need for society to encourage actors to cooperate without trust (risk-taking). How can we as a society build institutions to encourage initial cooperation without trust, so that mutually trustful cooperative relationships can eventually emerge, which the society will then not have to monitor or police? Second, our findings suggest that, to the extent that individuals in natural settings can separate trust from cooperation in order to eventually achieve high levels of trustful cooperation, social psychologists' heavy use of and reliance on PDs as a model of mixed-motive game might not be appropriate and might actually underestimate the extent and possibility of cooperation in society.

Acknowledgements

The experiments reported in this article were supported by grants to the first author from the Japan Society for the Promotion of Science. He is grateful to the Center for Advanced Study in the Behavioral Sciences for providing an opportunity to prepare this manuscript.

NOTES

1. At each subsequent shift, the payoffs change by 10% of the original matrix – the central matrix in Figure 1 – not 10% of the current matrix.

2. In this matrix, each player's dominant choice is not to return (because T > R and P > S). The total welfare (i.e. the sum of the two players' payoffs) is greater when both return than when neither returns (Pareto-superiority of mutual cooperation to mutual defection). These features correspond to the defining elements of a PD game.
3. The average cooperation rate and the *F*-value reported here are slightly different from those reported in the original study since we include all subjects in our analysis here whereas some subjects who indicated lack of understanding of the experiment were excluded from the original analysis. The values reported here are more conservative regarding the effect of the game. Figure 2 is based on Matsuda and Yamagishi's figure.
4. Only the Random PD/D*c* condition was newly added to the Japanese data.

REFERENCES

Arneson, R. J. 1982. 'The Principle of Fairness and Free-rider Problem.' *Ethics* 92: 616–33.

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Berg, J., J. Dickhaut and K. McCabe. 1995. 'Trust, Reciprocity, and Social History.' *Games & Economic Behavior* 10: 122–42.

Bolton, G. E., E. Katok and A. Ockenfels. 2003. 'How Effective Are Electronic Reputation Mechanisms? An Experimental Investigation.' Paper presented at the First International Symposium on Online Reputation, Sloan School of Management, Massachussets Institute of Technology, Boston, April 26–37.

Braithwaite, V. and M. Levi, eds. 1998. *Trust and Governance*. New York: Russell Sage Foundation.

Cook, K. S., ed. 2001. *Trust in Society*. New York: Russell Sage Foundation.

Cook, K. S., T. Yamagishi, C. Cheshire, R. Cooper, M. Matsuda and R. Mashima. 2005. 'Trust Building via Risk Taking: A Cross-Societal Experiment.' *Social Psychology Quarterly* 68: 121–42.

Dasgupta, P. 1988. 'Trust as a Commodity.' In *Trust*: *Making and Breaking Cooperative Relations*, ed. D. Gambetta, pp. 49–72. Oxford: Blackwell.

Fukuyama, F. 1995. *Trust*: *The Social Virtues and the Creation of Prosperity*. New York: Free Press.

Gambetta, D., ed. 1988. *Trust*: *Making and Breaking Cooperative Relations*. Oxford: Basil Blackwell.

Hardin, R. 2002. *Trust and Trustworthiness*. New York: Russell Sage Foundation.

Hayashi, N., E. Ostrom, J. Walker and T. Yamagishi. 1999. 'Reciprocity, Trust, and the Sense of Control: A Cross-societal Study.' *Rationality and Society* 11: 27–46.

Hofstede, G. H. 1980. *Culture's Consequences*: *International Differences in Work-related Values*. Beverly Hills, CA: Sage Publications.

Kakiuchi, R. and T. Yamagishi. 1997. 'General Trust and the Dilemma of Variable Interdependency.' *Research in Social Psychology* 12: 212–21. (In Japanese with an English abstract.)

Kiyonari, T., S. Tanida and T. Yamagishi. 2000. 'Social Exchange and Reciprocity: Confusion or Heuristic?' *Evolution and Human Behavior* 21: 411–27.

Knack, S. and P. Keefer. 1997. 'Does Social Capital Have an Economic Payoff? A Cross-country Investigation.' *Quarterly Journal of Economics* 112: 1251–88.

Kollock, P. 1993. 'Cooperation in an Uncertain World: An Experimental Study.' *Sociological Theory and Methods* 8: 3–18.

Kramer, R. M. 1999. 'Trust and Distrust in Organizations: Emerging Perspectives, Enduring Questions.' *Annual Review of Psychology* 50: 569–98.

Kreps, D. 1990. 'Corporate Structure and Economic Theory.' In *Perspective on Positive Political Economy*, eds. J. Alt and K. Shepsle, pp. 90–143. Cambridge: Cambridge University Press.

Lawler, E. J. 1986. 'Bilateral Deterrence and Conflict Spiral: A Theoretical Analysis.' In *Advances in Group Processes, Vol. 3*, ed. E. J. Lawler, pp. 107–30. Greenwich, CT: JAI Press.

Lawler, E. J. and J. Yoon. 1993. 'Power and the Emergence of Commitment Behavior in Negotiated Exchange.' *American Sociological Review* 58: 465–81.

Lawler, E. J. and J. Yoon. 1996. 'Commitment in Exchange Relations: Test of a Theory of Relational Cohesion.' *American Sociological Review* 61: 89–108.

Lindskold, S. 1978. 'Trust Development, the GRIT Proposal, and the Effects of Conciliatory Acts on Conflict and Cooperation.' *Psychological Bulletin* 85: 772–93.

McKelvey, R. D. and Palfrey, T. R. 1992. 'An experimental study of the centipede game.' *Econometrica* 60: 803–36.

Macy, M. W. 2002. 'Review of Trust in Society.' *Contemporary Sociology* 31: 473–5.

Mashima, R., T. Yamagishi and M. Macy. 2004. 'Trust and Cooperation: A Comparison of Ingroup Preference and Trust Behavior between American and Japanese Students.' *The Japanese Journal of Psychology* 75: 308–15. (In Japanese with an English abstract.)

Matsuda, M. and T. Yamagishi. 2001. 'Trust and Cooperation: An Experimental Study of PD with Choice of Dependence.' *Japanese Journal of Psychology* 72: 413–21. (In Japanese with an English abstract.)

Osgood, C. E. 1962. *An Alternative to War or Surrender*. Urbana: University of Illinois Press.

Pruitt, D. G. and M. J. Kimmel. 1977. 'Twenty Years of Experimental Gaming: Critique, Synthesis, and Suggestions for the Future.' *Annual Review of Psychology* 28: 363–92.

Putnam, R. D. 1993. *Making Democracy Work*: *Civil Traditions in Modern Italy*. Princeton, NJ: Princeton University Press.

Rosenthal, R. 1981. 'Games of Perfect Information, Predatory Pricing, and the Chain Store Paradox.' *Journal of Economic Theory* 25: 92–100.

Terai, S., Y. Morita and T. Yamagishi. 2003a. 'Trust and Assurance in Ongoing Relations: An Experimental Study Using the Prisoner's Dilemma with Variable Dependence.' *Research in Social Psychology* 18: 172–179. (In Japanese with an English abstract.)

Terai, S., Y. Morita and T. Yamagishi. 2003b. 'Trusting Behavior and Cooperative Relation in the Selective Play Situation: An Experimental Study Using the Prisoner's Dilemma with Variable Dependence.' Center for the Study of Cultural and Ecological Foundations of Mind Working Paper Series, No. 14. (In Japanese with an English abstract.)

Thibaut, J. W. and H. H. Kelley. 1959. *The Social Psychology of Groups*. New York: Wiley.

Van Lange, P. A. M. and K. Visser. 1999. 'Locomotion in Social Dilemma: How People Adapt to Cooperative, Tit-for-tat, and Noncooperative Partners.' *Journal of Personality and Social Psychology* 77: 762–773.

Yamagishi, T. 1998. *The Structure of Trust*: *The Evolutionary Game of Mind and Society*. Tokyo: University of Tokyo Press. (In Japanese, English translation available at http://lynx.let.hokudai.ac.jp/members/yamagishi/)

Yamagishi, T., K. S. Cook and M. Watabe. 1998. 'Uncertainty, Trust and Commitment Formation in the United States and Japan.' *American Journal of Sociology* 104: 165–194.

Yamagishi, T. and R. Kakiuchi. 2000. 'It Takes Venturing into a Tiger's Cave to Steal a Baby Tiger: Experiments on the Development of Trust Relationships.' In *The Management of Durable Relations*, ed. W. Raub and J. Weesie, pp. 121–3. Amsterdam: Thela Thesis Publishers.

Yamagishi, T. and K. Sato. 1986. 'Motivational Bases of the Public Goods Problem.' *Journal of Personality and Social Psychology* 50: 67–73.

Yamagishi, T. and M. Yamagishi. 1994. 'Trust and Commitment in the United States and Japan.' *Motivation and Emotion* 18: 129–166.

TOSHIO YAMAGISHI is Professor of Social Psychology in the Graduate School of Letters, Hokkaido University, Japan. His current research interests include co-evolution of social institutions and adaptive psychological mechanisms, and adaptive bases of psychological mechanisms that sustain a system of generalized exchange such as first-order and second-order punishment and in-group favoring behavior.

ADDRESS: Graduate School of Letters, Hokkaido University, N10 W7 Kita-ku, Sapporo, Japan 060–0810
[email: toshio@let.hokudai.ac.jp]

SATOSHI KANAZAWA is Lecturer in Management and Research Methodology at the London School of Economics and Political Science. His work on evolutionary psychology has appeared in peer-reviewed journals in all social sciences (psychology, sociology, economics, political science and anthropology) and in biology.

RIE MASHIMA is a doctoral student in the Graduate School of Letters, Hokkaido University. She is conducting both theoretical and experimental studies of strategies that can generate and sustain a system of generalized exchange.

SHIGERU TERAI is a post-doctoral research fellow in the Graduate School of Letters, Hokkaido University. He has just completed a PhD dissertation on cooperation and punishment between American and Japanese groups.