

Incremental approaches to establishing trust

Robert Kurzban · Mary L. Rigdon · Bart J. Wilson

Received: 30 January 2006 / Revised: 26 May 2007 / Accepted: 30 May 2007 /
Published online: 29 September 2007
© Economic Science Association 2007

Abstract We investigate cooperation using an incremental investment game in which the first-mover has the ability to make small, but increasing incremental investments in their counterpart. Our experiment is designed to test whether establishing trust in small increments is more effective than alternatives, including a one-shot investment game, a decrease only condition where the amount the first-mover sends to the second-mover must be less than the amount previously sent, and an unrestricted condition where the first-mover is not restricted by the amount previously sent. Although results were mixed, broadly, iteration affords greater cooperation than one-shot games and, when given the choice, participants seem to prefer to build trust gradually. Implications for institutional design are discussed.

Keywords Investment game · Trust · Reciprocity · Bargaining · Cooperation · Experimental economics

JEL Classification C72 · C91

R. Kurzban
Department of Psychology, University of Pennsylvania, 3720 Walnut St., Philadelphia, USA
e-mail: kurzban@psych.upenn.edu

M.L. Rigdon (✉)
Research Center for Group Dynamics, Institute for Social Research, The University of Michigan,
426 Thompson St., 48106 Ann Arbor, USA
e-mail: mrigdon@umich.edu

B.J. Wilson
Interdisciplinary Center for Economic Science, George Mason University, 4400 University Drive,
MSN 1B2, 22030 Fairfax, USA
e-mail: bwilson3@gmu.edu

1 Introduction

Consider two agents who can make a series of sequential mutually profitable exchanges. One party (the “buyer”) can defect during the exchange, failing to compensate the other party (the “seller”) who provides an initial investment. If we assume financially motivated self-interested actors, then the second party to act will defect when the incentives to do so are sufficiently great, so the first party should not invest. This inefficient outcome in which gains in trade are left unrealized has been dubbed the “hold-up problem”. Pitchford and Snyder (2004) show that a solution to the problem is to have the seller extend a large initial investment at the beginning of the iterated set of interactions, followed by successively smaller amounts. This induces the buyer to reciprocate so the seller then will continue with the iterated interaction.

This analysis holds under canonical assumptions regarding pure self-regarding preferences. However, it is reasonable, in the face of increasing evidence, to suppose that there might be behavioral mechanisms that extend beyond this narrow set of preferences (Camerer 2003). For example, one can imagine a very simple preference to risk moving first in an exchange that increases as the history of successful exchanges lengthens. A basic form of such a preference would be “Tit-for-Tat”, which chooses to cooperate with past cooperative partners but not with past defecting partners (Axelrod 1984). A more complex variation of this, also from evolutionary biology, is a strategy called “Raise the Stakes” (RTS), which increases its investment with partners who have played sufficiently reciprocally in the past round in a dynamic environment (Roberts and Sheratt 1998). This strategy did extremely well under the parameters of their simulations (Friedman and Hammerstein 1991). The success of such strategies implies that people might have risk preferences in exchanges such that they are willing to take relatively small risks initially, increasing their risks as a relationship develops. Indeed, precisely this sort of behavioral strategy has been shown to be effective (Blonski and Probst 2004; Pillutla et al. 2003; Rempel et al. 1985; Watson 2002). The prescription has been applied to interactions ranging from interpersonal (economic) relationships (Lewicki and Bunker 1996) to international politics (Lindskold 1978; Osgood 1962; Schelling 1960).

Consider the implications for agents with two very different types of preferences. With self-regarding preferences, the Pitchford and Snyder solution of progressively smaller investments should yield social welfare gains.¹ If, instead, agents have preferences to extend trust as a positive function of a history of past successful transactions, the ability to engage in progressively greater investments should yield social welfare gains. Note that the reverse is true as well—agents who prefer to risk investment *only* when there is a history of successful transactions are unlikely to risk large initial transactions, by definition.

The focus here is on comparing the efficiency of different institutions using incremental variants of the Investment Game (Berg et al. 1995). We examine mechanisms that force progressively increasing transactions, mechanisms that force progressively decreasing transactions, and two control conditions. In one such condition we simply

¹Note this argument turns on self-regarding agents with the ability to perform computations associated with backward induction.

ask, if first movers are unrestricted in their investments over time, how do they allocate those investments in second movers? We find that, while there were some differences between our populations, iterated investment games yielded greater efficiency than single-shot games and, more interestingly, when iterations were unrestricted, there was a preference for starting with smaller levels of investment and increasing it, rather than the reverse. These results are highly suggestive. How trust is accorded to others over time is important for the question of how to design institutions that allow social welfare gains to be captured.

2 Prior research

The environments we will be concerned with will be variants of the two-person Investment Game or trust game (Berg et al. 1995). In this game Player 1 (P1) and Player 2 (P2) are allocated endowments X_0 and Y_0 (respectively). P1 has the opportunity to send some, all, or none (X), of X_0 to P2. Any amount sent is multiplied by some factor $r > 0$. P2 then has the decision of how much of rX (Y) to send back to P1 and how much to keep for herself. Trust is thus operationalized by the amount P1 sends to P2, and trustworthiness is operationalized by the fraction returned (i.e., amount returned divided by rX).

In one-shot, double-blind experiments with anonymous randomly matched subjects ($X_0 = Y_0 = \$10$ and $r = 3$), Berg et al. (1995) found that 30 of 32 P1's sent money to P2's, with $X = \$5.16$ sent on average, and transfers of $X = \$5$ had an average payback of $Y = \$7.17$. These results suggest not only that people do trust others in the one-shot game, but also that trust is reciprocated.²

Roberts and Renwick (2003) had participants play a series of games resembling the Prisoner's Dilemma (PD) in which investment could be varied from one round to the next. They found that investment indeed increased over the course of successive rounds, suggesting that players used a strategy of gradually increasing their trust if previous investment in their partner was reciprocated.

Andreoni and Samuelson (2006) conducted experiments similar in spirit and broadly are in accord with these findings. They used a two period PD game in which the amount of money at risk was constant in the two periods, but the distribution was allowed to vary across time periods. Their results indicated that when more of the payoffs were concentrated in the second period, the amount of cooperation was largest. Therefore, players were best off when they were able to "start small", allowing the second period to become relatively more important.

There are, however, data from the laboratory suggesting that gradual increases in trust are not always effective. In a one-shot trust game, Pillutla et al. (2003) demonstrate that "precipitous trusting"—showing a large amount of trust in a one-shot interaction—does indeed lead to more reciprocation than showing intermediate

²A replication study by Ortmann et al. (2000) found similar results. See Camerer (2003) for a summary and discussion of some results of experiments testing robustness. There are also results from a binary-choice trust game demonstrating that experimental subjects exhibit significant amounts of trust and reciprocity (McCabe et al. 2002, 2003; Eckel and Wilson 2003).

amounts of trust. They conclude that their results “stand in marked contrast to the traditional prescriptions for trust development, suggesting that attributional processes might facilitate all-or-none rather than gradual trust development. . . ” (p. 454).³

Another interesting example is behavior in the centipede game—a game that gives two players sequential opportunities to defect, with the gains from trade increasing over time if they choose instead to cooperate. When the number of players is increased from two to three and stakes are sufficiently high, results tend to yield the equilibrium predicted by standard theory: defection occurs very early in the interaction, with potential gains from trade left on the table (Rapoport et al. 2003). Similar results consistent with this view of decision-making can be seen in end-game effects, in which agents do indeed confiscate gains if they have knowledge that there will be no future interactions (Slonim et al. 2001; Bolton et al. 2003). Whereas the kind of model Pitchford and Snyder propose predicts these end-game effects, models suggesting that trust is consistently built over time imply that these effects should be relatively rare, as the end of the game is exactly when the maximum amount of trust has been built.

How trust is accorded to others is important for the question of how to design institutions that allow social welfare gains to be captured (Zak and Knack 2001; Dufwenberg and Kirchsteiger 2004; Falk and Fischbacher 2006). Andreoni and Samuelson’s (2006) work takes a step toward answering the question about how to “structure interactions so as to enhance economic efficiency by fostering cooperation”, (p. 1). Kurzban et al. (2001) also address how a particular institutional arrangement in the context of a public goods game can reverse the usual trend for contributions to public goods to decrease over the course of repeated play (see also Dorsey 1992). Their results have recently been replicated in Japan (Ishii and Kurzban 2007, *in press*), suggesting that the institutional mechanism is robust to the cultural differences between the West and East (Markus and Kitayama 1991; Triandis 1995).

Gale (1995, 2001) introduced the “monotone game” in which players must choose non-decreasing stage-game strategies over time. A monotone game is a dynamic implementation of a one-shot game. The difference is that players gradually commit themselves in Gale’s monotone game as opposed to first engaging in a large investment as proposed by Pitchford and Snyder. Choi et al. (2006) experimentally applied the theory of a monotone game to a voluntary contribution mechanism (VCM) for public goods and found that the dynamic implementation of the one-shot game leads to higher contributions than a one-shot game. Duffy et al. (2006) similarly conducted a laboratory experiment on a dynamic VCM but one based upon the model of Marx and Matthews (2000). The Marx and Matthews environment included a discrete positive benefit for completing the public good. Duffy et al. also found greater contributions in a dynamic versus a single-shot game.

More generally, the search for social-welfare improving institutions has been the subject of a certain amount of research attention for some time (North 1981, 2005). There is, however, continued debate about the way in which trust is built over time.

³Weber et al. (2004) suggested that: “Precipitous trusting acts that benefit the trusted party will . . . increase the likelihood of reciprocity” and “accelerate the development of mutual trust” (p. 89).

3 Experimental treatments

While one-shot games are valuable for asking certain kinds of questions about behavior and preferences, as Weber et al. (2004) note: “trust development is almost necessarily sequential, moving back and forth between the two parties via turn taking and reciprocity” (p. 79). Our current interest is precisely in this “back and forth” of iterated interactions. The behavioral mechanisms described above suggest two very different dynamics for trust-based interactions over time, leading to two focal questions. First, do people use the gradual building of reciprocated trust over time to generate Pareto improving outcomes? Or, do they use the solution to the hold-up problem proposed by Pitchford and Snyder, which would be effective if people bring to bear rational expectations and have self-regarding preferences? Second, what kind of institutional mechanism will be most effective in generating trust and trustworthiness? To address these questions we use variants of the Investment Game described above allowing repeated exchange with an endogenous end-point and vary the constraints on P1’s subsequent investment decisions.

If the behavioral mechanisms that make the hold-up problem a problem—i.e., self-regarding preferences—then the worst case is a single-shot interaction. In such an exchange, with no subsequent possibilities for exchange possible, sellers should never sell. In iterated interactions, transactions in which exchanges start small and get larger should be particularly unsuccessful by the same logic. It follows that if agents are self-regarding and can reason about the hold-up problem, then, when given the choice, they should prefer to start their transactions large and decrease them over subsequent rounds. If these predictions are not borne out by empirical data, then it is reasonable to question the assumptions—including the postulated behavioral mechanisms—that underpin the logic of the hold-up problem.

As with other investment games, in our experiment, both P1 and P2 receive a \$12 endowment.⁴ P1 is able to send any amount of his endowment to P2 in \$0.25 increments. Any amount sent to P2 is tripled. P2 then decides how much money, if any, to return to P1 and how much to keep. This is the Baseline (B) treatment. There are three treatments we consider: Increase Only, Decrease Only, and Unrestricted. The experimental treatments add iteration to the canonical investment game environment, allowing players to interact over multiple periods. Under all of the mechanisms, a *round* is divided into at most 10 *periods*, with the end period able to be determined endogenously by the players. Within a particular round, players are paired with the *same person* for all the periods. In the Increase Only (IO) condition, P1 must send strictly *more* money to P2 than he sent in the previous period. This condition allows us to examine whether or not both players are better off when trusters begin by using small, increasing incremental investments. In the Decrease Only (DO) condition, P1 must send strictly *less* money to P2 than he sent in the previous period. This condition imposes a structure requiring the largest initial investment decision to be made in the first period by P1, followed by decreasing levels over time, as in the Pitchford and Snyder solution to the hold-up problem. In the Unrestricted (UR) condition,

⁴The \$12 is in experimental dollars and is converted to U.S. dollars at the exchange rate of 0.2 experimental dollars to US \$1.

P1's decision about an amount to send to P2 is *independent* of his previous decisions. This condition provides no constraints on the transfer decisions P1's can make. At the completion of a round, the players are re-paired, never meeting the same counterpart again.

4 Predictions

What can we predict will occur under the two different behavioral mechanisms described above? Assuming self-regarding preferences, the Decrease Only institution should generate substantial gains in trade because of the process described by Pitchford and Snyder because this kind of institution forces a solution to the hold-up problem. That is, self-regarding actors have an incentive to be trusting and be trustworthy due to possible future gains in trade available through this mechanism. In contrast, if agents prefer to trust *only* in large amounts only when there has been a history of successful trust interactions with the other person, then the institutional mechanism should be less successful.

Next, consider the Increase Only mechanism. If our subjects have self-regarding preferences, they should be unwilling to trust in the last move and, by backward induction, in early rounds as well. In contrast, if subjects prefer to build trust over time, and will trust in large amounts if there has been a history of fulfilled trust, then this mechanism should show substantial realization of the gains available from trade.

Finally, the Unrestricted condition should illustrate preferences when they are allowed iterated interaction in the trust game. If models that suggest people prefer to incrementally increase their level of trust are correct, behavior in the Unrestricted case should resemble the Increase Only condition. If people have self-regarding preferences, and expect this of others, then behavior in the Unrestricted case should resemble Pitchford and Snyder's model because agents can solve the hold-up problem by starting with large amounts of trust and leaving additional gains in trade in the future as the incentive to reciprocate trusting moves.

In short, the Decrease Only mechanism will yield Pareto-superior outcomes relative to the Nash equilibrium if the relevant behavioral mechanism subjects bring to bear are rational self-regarding preferences, but the Decrease Only mechanism will yield Pareto-superior outcomes if the relevant behavioral mechanism subjects bring to bear are preferences to trust a little at the beginning and a great deal if and only if there is a history of successful transactions. The Unrestricted case should illustrate what subjects choose to do when left to their own devices.

5 Methods

Sessions of the experiment were conducted with undergraduate students from a variety of majors at both George Mason University (Mason) and the University of Pennsylvania (Penn). Three sessions of each treatment were completed at Mason and two sessions of each treatment were completed at Penn, with a total of five sessions for each treatment condition. A session consisted of 12 subjects resulting in a total of 60

subjects per treatment, with half playing the role of P1 and the other half playing the role of P2.

A participant was paid \$5 for showing up on time, and immediately seated at a computer terminal. The instructions were computerized and self-paced. Subjects were told that the experiment would last a total of 5 rounds. It was common information that they would be re-paired each round, never meeting the same counterpart again. In the treatment conditions, it was explained that a *round* would last at most 10 *periods*. In these periods, players were paired with the *same* individual. All players were able to end the round at any time by clicking on the “Terminate Round” button when prompted to make a decision. Figures 1 and 2 present the computer screens that P1 and P2 saw, respectively. Total earnings were indicated for P1 on the left hand side of the screen and for P2 on the right hand side. There were also bars graphically displaying the \$12 endowments for each player. Below the bar indicating one’s own earnings was a box where players entered the transfer amount; players then clicked the “Send” button and were prompted to either confirm their decision or change it. The particular example in the figures is for the IO treatment. Figure 1 shows Round 1, Period 1, prior to P1 making a decision. Looking at Fig. 2, one can see that P1 transferred \$5 to P2 in Period 1. The bar of \$15 represents the amount transferred times the multiplier. In this example, P2 is deciding to return \$7.25 out of a possible \$15. P1 now had a decision to make (see the right hand side of Fig. 1): either terminate the round, in which case P1 would earn \$14.25 and P2 would earn \$19.75, or transfer an amount to P2 strictly larger than the \$5 amount transferred in Period 1 and less than or equal to the remaining endowment of \$7. Looking at Fig. 2, one can see that P1 transferred \$5.25 to P2 in Period 2, and P2 returned \$8 out of a possible \$15.75. P1 must terminate the round because he only has \$1.75 left of his endowment. Therefore, in Round 1, P1 and P2 earned, respectively, \$17 and \$27.50.

As in the original Investment Game experiment, ours was run double-blind so that no one, including the experimenter, could map decisions to the identity of the decision maker. We implemented the double-blind procedure by having each subject prior to the beginning of the session enter a code into their machine that they select and write it down on a slip of paper. At the completion of the experiment, each subject—privately and one at a time—slid an envelope containing the slip of paper with their chosen code under a closed door. A monitor inside the room looked up the code, put the dollar earnings in the envelope, and slid the envelope back under the door.⁵ The sessions took about 1 hour to complete.

6 Results

The standard logic of backward induction implies that there should be no trustworthy behavior (by P1) and, therefore, no trust (by P2). This is the case even in the treatment conditions in which interaction is repeated, but finite because of P1’s budget constraint of \$12. Further, the reasoning suggested by Pitchford and Snyder (2004) implies that the mechanism by which trust begins high and decreases over time

⁵A slightly different procedure was used at Penn, though similarly preserving anonymity.

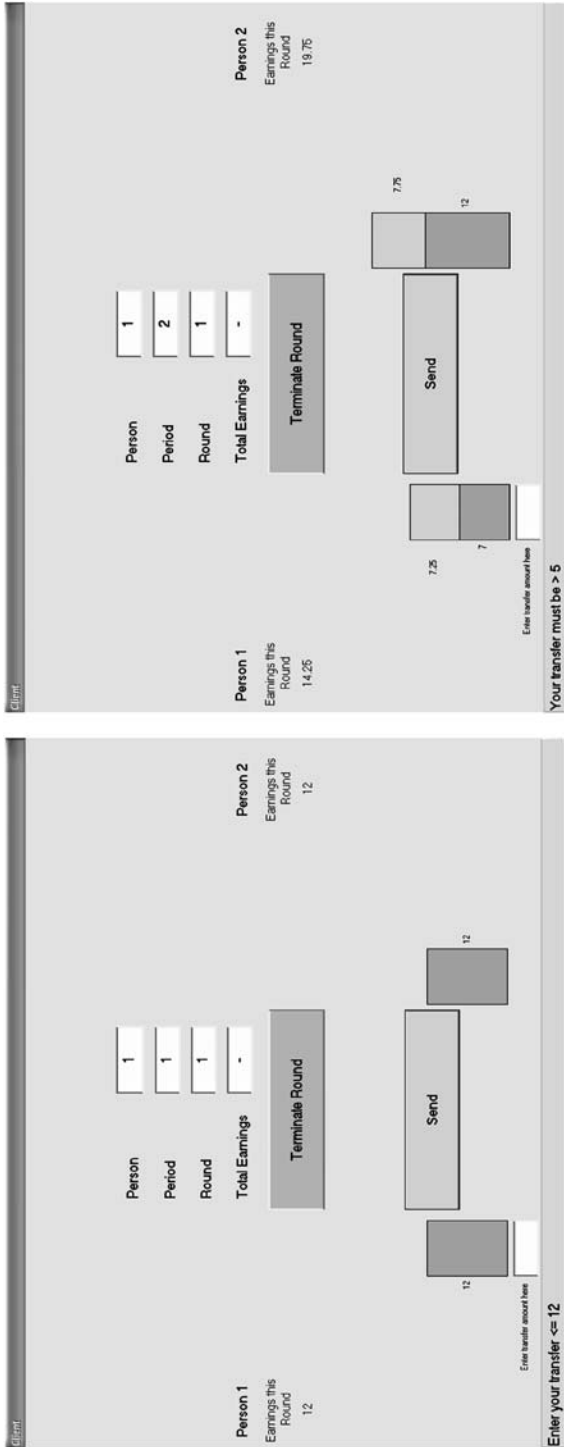


Fig. 1 Player 1's screen

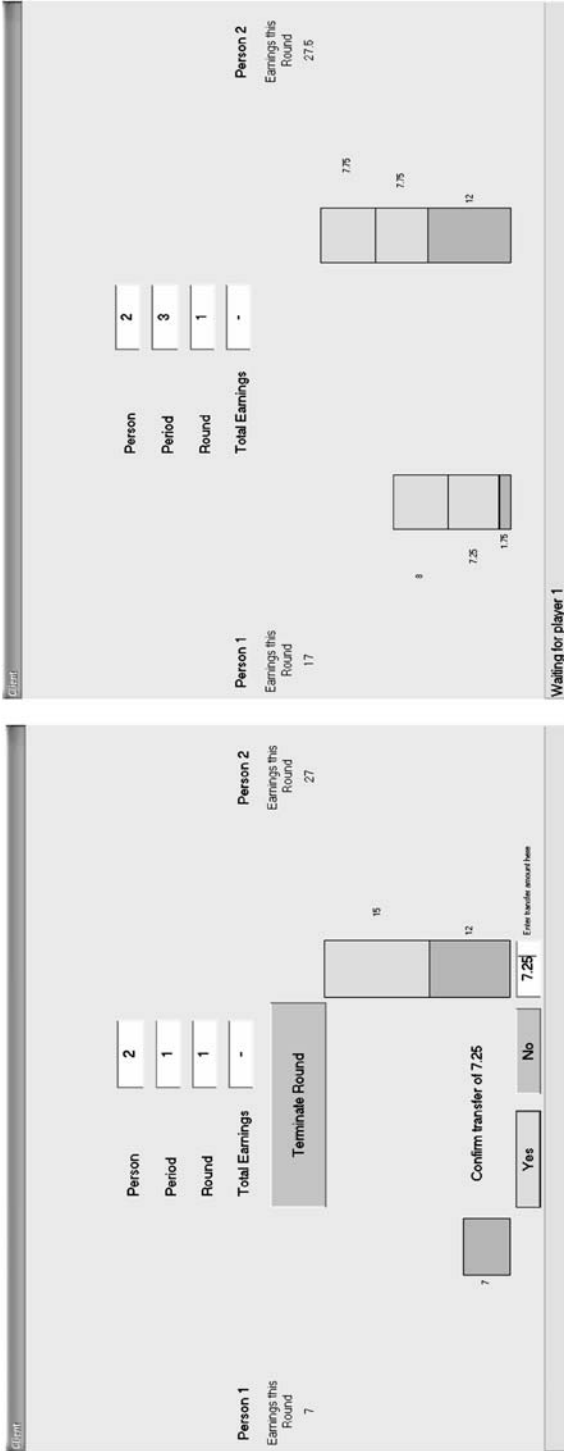


Fig. 2 Player 2's screen

Table 1 Summary statistics across condition by university

		Condition			
		Baseline	Unrestricted	Increase only	Decrease only
Transfer (<i>X</i>)	Penn	4.63 (4.14)	8.43 (3.83)	8.75 (3.49)	7.17 (3.68)
	Mason	6.01 (3.85)	9.57 (2.97)	8.02 (3.74)	8.18 (4.12)
<i>Y</i> / <i>X</i>	Penn	1.47 (.89)	1.47 (.67)	1.55 (.81)	1.47 (.71)
	Mason	1.04 (.75)	1.57 (.60)	1.22 (.70)	1.50 (.72)
P1 Earnings	Penn	14.1 ^a	16.3 ^{ab}	16.9 ^b	15.1 ^{ab}
	Mason	12.0 ^a	18.0 ^b	14.3 ^c	16.7 ^b
<i>Y</i> / <i>3X</i>	Penn	0.48 ^a	0.48 ^a	0.48 ^a	0.49 ^a
	Mason	0.35 ^a	0.52 ^b	0.41 ^c	0.50 ^b

Mean (standard deviation). *X* is the amount sent by P1 and *Y* is the amount returned by P2. Cells in each row sharing a superscript do not differ significantly from one another; cells that differ do so at $p < .05$ (two-tailed test).

should generate more trustworthy and trusting behavior. These predictions are the background against which we pose the following six questions. Descriptive statistics for P1’s Transfer (*X*) and the ratio of the amount returned from P2 divided by the amount transferred (*Y*/*X*) are given in Table 1.

Question 1 Do the Incremental mechanisms produce more trust than Baseline on the part of P1’s?

The quantitative results are derived by analyzing the data with a linear mixed-effects model for repeated measures. The treatment effects (*IO*, *DO*, and *UR*) and population effect (*Mason*) and the interaction effects of the treatments and population are modeled as zero-one fixed effects, while the sixteen independent sessions and the six P1’s within each session are modeled as random effects, e_i and ζ_{ij} , respectively. Specifically, we estimate the model

$$\begin{aligned}
 PITransfers_{ijr} = & \mu + e_i + \zeta_{ij} + \beta_1 UR_i + \beta_2 IO_i + \beta_3 DO_i + \beta_4 Mason_i \\
 & + \beta_5 UR_i \times Mason_i + \beta_6 IO_i \times Mason_i + \beta_7 DO_i \times Mason_i + \varepsilon_{ijr}
 \end{aligned}$$

where $e_i \sim N(0, \sigma_1^2)$, $\zeta_{ij} \sim N(0, \sigma_2^2)$, and $\varepsilon_{ijr} \sim N(0, \sigma_3^2)$. The sessions are indexed by i ; the P1-subjects within each session are indexed by $j = 1, \dots, 6$; and the repeated rounds are indexed by $r = 1, \dots, 5$. The dependent variable $PITransfers_{ijr}$ is the amount transferred by subject j in session i in round r .

The second column in Table 2 reports the regression results; see regression (1). Transfers significantly differ from 0 across all conditions, falsifying the subgame

Table 2 Estimates of linear mixed-effects models for: (1) Transfers by Person 1 and (2) Amount Returned by Person 2

	(1) Estimate	(2) Estimate
μ	4.63**** (0.82)	-0.27 (0.80)
Unrestricted	3.78*** (1.15)	0.67 (1.52)
Increase Only	4.12*** (1.15)	1.83 (1.61)
Decrease Only	2.43* (1.15)	0.69 (1.37)
Mason	1.39 (1.05)	0.98 (1.13)
Unrestricted \times Mason	-0.22 (1.49)	-2.04 (2.24)
Increase Only \times Mason	-2.12 (1.49)	-2.17 (2.07)
Decrease Only \times Mason	-0.26 (1.49)	-1.77 (1.86)
Transfer	—	1.54*** (0.22)
Transfer \times Unrestricted	—	-0.15 (0.32)
Transfer \times Increase Only	—	-0.19 (0.32)
Transfer \times Decrease Only	—	-0.20 (0.32)
Mason \times Transfer	—	-0.66* (0.29)
Mason \times Transfer \times Unrestricted	—	0.93* (0.43)
Mason \times Transfer \times Increase Only	—	0.53 (0.42)
Mason \times Transfer \times Decrease Only	—	0.89* (0.41)

$n = 600$. The linear mixed effects model for repeated measures treats each session as one degree of freedom with respect to the treatments. * = $p < .10$, ** = $p < .05$, *** = $p < .01$, **** = $p < .001$; standard errors in parentheses

perfect equilibrium prediction, as has been found in other experimental trust games (Bohnet and Zeckhauser 2004; Croson and Buchan 1999; McCabe et al. 2003; Ortmann et al. 2000). In our Baseline condition, P1's send $\hat{\mu} = 4.63$ (out of a maxi-

mum of 12), which is consistent with prior work. Further, in each of the incremental treatment conditions in which transfers are (potentially) iterated among pairs, more money is transferred by P1 than in the Baseline condition. In the IO treatment, P1's send $\hat{\mu} + \hat{\beta}_2 = 8.75$ which is highly significant. Similarly, P1's in the DO and UR send $\hat{\mu} + \hat{\beta}_3 = 7.06$ and $\hat{\mu} + \hat{\beta}_1 = 8.41$, respectively. The effect of the treatment condition does not vary by participant population as all coefficients involving *Mason* are highly insignificant. In general, then, iteration does increase transfers by P1's.

Question 2 Among the Incremental mechanisms, does the type of mechanism change the amount of trust (surplus) generated?

This issue is of particular relevance because the answer can shed light on how individuals in two-person exchange attempt to establish a reciprocal relationship. Whereas the IO mechanism might be predicted to facilitate small but increasing amounts of trust and trustworthiness, the DO mechanism should be of particular value for overcoming the hold-up problem, if that is preventing trust in these interactions. With a Likelihood Ratio test for the model reported in column 2 of Table 2, we fail to reject the null hypothesis that there is no difference in the additional amounts sent by P1's in the IO, DO, and UR mechanisms ($p = .20$).

Question 3 Do P1's earn more in the incremental conditions?

Holding aside interactions with subject population (see Question 4 below), the proportion transferred back from P2 to P1 does not vary significantly across treatment conditions. Given that trust is greater in the incremental conditions, one might therefore predict that P1's earned more in these conditions. A 2 (University) \times 4 (Treatments) ANOVA on earnings for P1 averaged across Rounds reveals a main effect for Treatment ($F(3, 112) = 12.66, p < .01$), no significant effect for University ($F(1, 112) = 0.37, p = .54$), and a significant interaction ($F(3, 112) = 4.83, p < .01$).

The significant interaction of University and Treatment suggests that our two populations are responding differentially to the treatment conditions. A one-way ANOVA of the Penn data reveals no significant effect at a 5% level of confidence ($F(3, 44) = 2.27, p = .09$). However, a similar one-way ANOVA of the Mason data is highly significant ($F(3, 68) = 15.26, p < .01$). Table 1 indicates the means and differences in P1's earnings across conditions. In short, P1's at Mason in the Unrestricted and Decrease Only conditions fare better than those in the Increase Only condition, but P1's in the IO condition earn more than those in the Baseline condition.

Question 4 Do the Incremental mechanisms generate more trustworthiness than the Baseline (one-shot) mechanism?

As discussed above, trustworthiness is quantified by P2's behavior. To answer this question, we expand the linear mixed effects model from above by including an additional fixed effect on a variable *Transfer*, which is the amount originally sent by P1 to

P2 before it was tripled. The dependent variable is the amount returned by P2 to P1. We also include two-term interaction effects of this variable with the UR, IO, DO, and Mason, as well as three-term interaction effects (e.g., $Mason \times Transfer \times UR$). *Transfer* measures the *proportion* (rather than absolute *amount*) returned.⁶ The third column of Table 2 reports the regression; see regression (2). The highly significant coefficient on *Transfer* shows that P2's return 1.54 of the amount that they are sent, i.e. P1's receive a 54% return on the investment. This rate of return is not affected by the UR, IO, and DO treatments. Hence, in light of our answer to question 1, the welfare of the P1's is only enhanced in the incremental treatments by the additional amounts that P1's are willing to invest in P2's.

This effect, however, is qualified by a statistically significant interaction with subject population. The bottom half of Table 2 (column 3) shows that P2's at Mason returned a smaller proportion in the Baseline and IO condition. P2's at Mason keep these gains; in particular in the Baseline condition, whereas at Penn gains from trade are divided roughly evenly. Relative to Penn, P2's in the Baseline condition at Mason return $1.54 - 0.66 = 0.88$. The IO treatment effect at Mason when interacted with *Transfer* is not significantly different than the Baseline (the coefficient estimate of 0.53 for $Mason \times Transfer \times IO$ is statistically insignificant). This illustrates a possible problem in one-shot interactions. As addressed in more detail below, this can be construed as preliminary evidence of the effect of the hold-up problem or, more basically, end-game effects. The incremental UR and DO mechanisms, however, appear to solve this problem at Mason. In stark contrast, the estimated coefficients for $Mason \times Transfer \times UR$ and $Mason \times Transfer \times DO$ are statistically significant and positive such that they offset the Mason population effect observed for the Baseline and IO treatment. That is, P2's in the UR and DO treatment at Mason send back $1.66 (= 1.54 - 0.15 - 0.66 + 0.93)$ and $1.57 (= 1.54 - 0.20 - 0.66 + 0.89)$, respectively, of the amounts transferred to them by P1. Taken together, these population effects at Mason are a clear case of exceptions (Baseline and IO treatments) that appear to prove the rule—incremental mechanisms, and in particular the DO mechanism—can support trust and reciprocity where it might not exist otherwise.

Question 5 When transfers are unrestricted (UR), how do people use the Incremental mechanism?

The Unrestricted condition allows us to ask additional questions regarding how people use the incremental mechanism when the increase only and decrease only constraints are lifted. Broadly, how people behave across periods allows us to investigate whether people use initial large contributions to overcome the hold-up problem or build trust using small incremental investments.

Very generally, if players are concerned about the hold-up problem, it is reasonable to expect that they would transfer an intermediate amount of money in the first period of the interaction (since investments must gradually shrink over time). In contrast, if

⁶For example, a coefficient of 1.5 on *Transfer* indicates that P2 returned 1.5 of what P1 sent to P2 before it was tripled (or equivalently, half of $3X$).

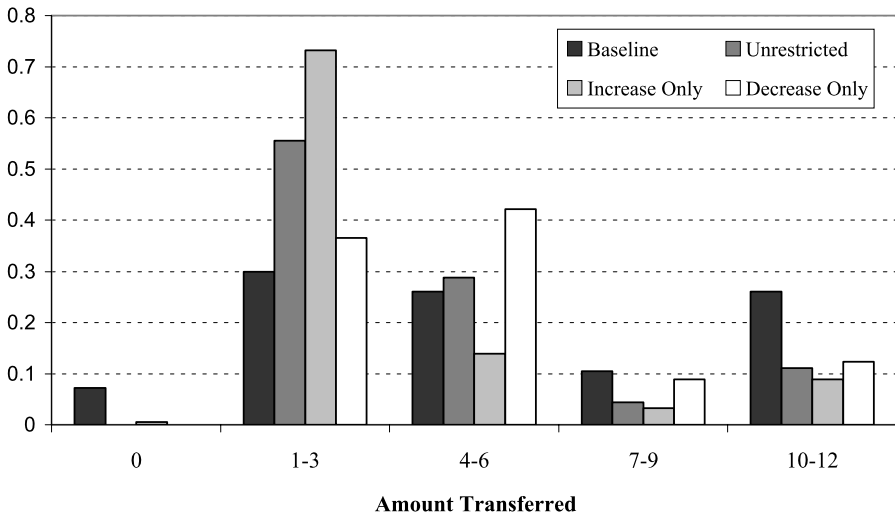


Fig. 3 Amount (\$) transferred by P1's by condition in the initial period, collapsed across round

players are trying to build trust incrementally, P1's can be expected to transfer relatively small amounts in the first period of the interaction. Further, if players engage in the type of reciprocity suggested by incrementally building trust, then transfers at time 2 should be related to time 1 such that if a trusting decision at time 1 is responded to by P2 positively (i.e., with P1 receiving at least as much as sent), then transfers at time 2 should be increased. The inverse should hold as well.

Figures 3 and 4 address these issues. As can be seen in Fig. 3, the result of first period contributions is somewhat equivocal: slightly more than half (55%) of contributions are between \$1 and \$3, the remainder of all first period contributions are \$4 or greater. However, it is worth pointing out that the distribution of initial transfers in the Unrestricted condition bears a certain resemblance to that in the DO condition. Indeed, Kolmogorov–Smirnov tests comparing the distributions of initial transfers in the Unrestricted condition indicates a significant difference with respect to the Increase Only condition ($p < .01$), but not the Decrease Only condition ($p = .35$).

Figure 4, however, is generally supportive of a model of increasing goodwill. Data from only Period 2 show that when P2's are untrustworthy (returning less than or equal to the amount sent by P1), P1's reduce subsequent transfers.⁷ In contrast, when P2's return more than the amount sent, P1's increase the amount transferred. Finally, in line with a general trend toward reciprocity, there is a strong and positive relationship between the amount sent and amount returned for all conditions: Baseline ($r = .65$), U ($r = .77$), IO ($r = .63$) and DO ($r = .74$).⁸

⁷Period 2 is the most informative period. It is the one for which we have the most data by virtue of the fact that not all pairs progressed beyond period 2.

⁸We also looked at whether the types of investments that P1's make differ between the top and bottom third performers in terms of total earnings. In all three treatments, we did not observe a difference between the amounts sent by P1's.

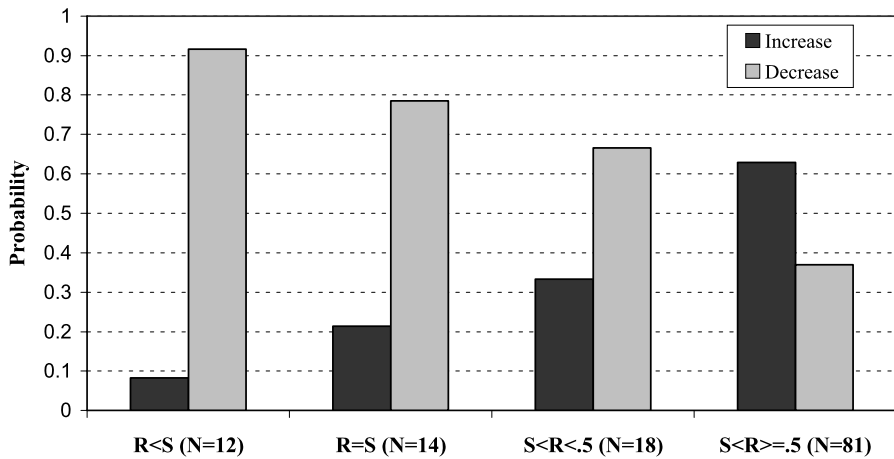


Fig. 4 Probabilities that P1 increased or decreased the amount sent in the second period conditional on P2's behavior in period 1. The *horizontal axis* refers to the relationship between the amount returned by P2 and the amount sent by P1. For example, " $S < R < .5$ " is the case in which P2 returned (R) more than was sent (S), but less than half of the gains from trade (.5)

Question 6 How does the ratio of gains from exchange vary by condition?

The ratio of the gains from exchange quantifies the extent to which surplus is divided relatively equally between the two participants involved in the exchange. A 2 (University) \times 4 (Treatments) ANOVA on the ratio of P2's transfer to P1's (Y) to the total value sent by P1 ($3X$), conditional on P1 having sent any positive amount, reveals a significant effect of Treatment ($F(3, 112) = 5.86, p < .01$), University ($F(1, 112) = 7.01, p < .01$), and a significant interaction ($F(3, 112) = 5.77, p < .01$). A separate ANOVA for each participant population indicates no significant difference among conditions at Penn ($F(3, 44) = .26, p = .85$), but a significant difference at Mason ($F(3, 67) = 15.29, p < .01$). Pair-wise differences of the ratio of gains from exchange ($Y/3X$) across treatments are shown in Table 1. At Mason, P1's in the Increase Only receive a greater fraction of the gains from exchange than the P1's in the Baseline. Furthermore, P1's in the Decrease Only and Unrestricted receive a greater fraction of the gains from exchange than the P1's in the Increase Only. In sum, P1's benefit from the incremental mechanisms at Mason given the observed opportunism of P2's in the Baseline treatment.

7 Discussion

We began by considering two possible behavioral mechanisms that could potentially be at work in transactions that occur over time. If people have rational, self-regarding preferences, then building trust slowly is doomed because players expect last-round defection, leading to the hold-up problem. If people, in contrast, prefer to extend trust (and reciprocate it) as a positive function of past successfully reciprocal transactions, then building trust slowly should be an effective way to reap welfare gains. To address

this issue, we examined how four different institutional regimes influence the decision to be trusting and trustworthy.

Broadly, in line with both predictions and previous work, each of the three iterated mechanisms increased the efficiency of the interactions between first- and second-movers in an investment game. Surplus was uniformly higher than the subgame perfect equilibrium prediction of no trust by P1's and the untrustworthiness of P2's. In all three iterated conditions, P1's cumulatively sent more to their P2 counterparts than in the one-shot baseline, and there was no difference among the conditions in the total amount sent by P1's. We did, however, observe a difference in P2 behavior across the universities. At Mason, P2's were less trustworthy than their Penn counterparts in the Baseline and Increase Only treatments, the two treatments in which P1's are most susceptible to the defecting, self-regarding behavior of P2's. In the Baseline treatment, P1's are susceptible to untrustworthy P2's because the interaction is one-shot. In the Increase Only treatment, there is a known ending point for P2's when P1 can no longer send anymore, and this final amount is the largest of any amount sent. It is under these conditions that the hold-up problem is most acute.

The fact that P1's send equivalent amounts across iterated conditions suggests that P1's expect P2's will be trustworthy even when P2's have an opportunity to profit from returning little or none of the money generated from P1's transfer. Even knowing that there is no shadow of the future that would provide an incentive for P2's to reciprocate, P1's expect P2's to be trustworthy.

At Penn, P1's are right to expect P2's to reciprocate: P2's reciprocate sufficiently in the Increase Only condition, for example, that P1's earn nearly \$7 more than their initial endowment. Not so at Mason, where P2's proved relatively untrustworthy in the Baseline and Increase Only treatments. Taken together, these results imply that (1) people are inclined to believe that trusting behavior will be reciprocated and (2) they are sometimes, but not always, wrong.

This finding, of course, mirrors findings in the one-shot version of the game, but carries additional information. For one thing, in cases in which exchanges are exogenously limited, it is clear that iteration in and of itself is not sufficient to guarantee realization of all the gains in trade. That is, repetition is clearly not a *sufficient* condition to solve the problems of untrustworthy behavior in the one-shot game. Iteration *can* improve outcomes, but this depends on the nature of the populations. This leads to the second conclusion, which is that, as others have suggested (e.g., Kurzban and Houser 2005; Kurzban et al. 2001), more attention ought to be paid to the diversity of types in populations and the inferences people draw from information they have about others. That is, it could be that with even a limited amount of information about their counterparts, people could make better judgments about when to be trusting (Frank et al. 1993). The heterogeneity we observe might be telling us that we ought to be paying more attention to partner selection, which looms larger as heterogeneity in trustworthiness among agents increases.

Finally, the results from the Unrestricted condition tell us something potentially important about (mean) preferences. When given the choice, people seem inclined toward building trust incrementally—starting relatively small with more than half of the participants offering \$1 to \$3 in the initial period and respond, at least in the second period, by increasing trust that was previously reciprocated, but not increasing

trust that was not reciprocated. This adds to previous evidence suggesting that people have a preference for trust to be built gradually (see above).

8 Caveats

The above discussion of between-University differences should not be taken to imply that we believe that we have identified critical population-level differences at the two institutions. With the limited samples here, we hesitate to draw conclusions about the populations *per se*. Rather, we can say that the effectiveness of the institutions differs in important ways depending on the participants. This is informative of the “strength” of the institution (see below).

Consider the canonical double-auction, which elicits similar behavior across very diverse populations (Smith 1991). That is, this institutional mechanism pulls behavior sufficiently strongly that individual differences are flattened. The present institutions are not as robust, a point which was difficult to know *ex ante*.

Finally, an important variable is whether the end point of interactions is known. Methodological constraints necessitated known, exogenously imposed endpoints in the present case, but there are no doubt many real-world contexts in which the endpoint is unknown and/or determined endogenously. Indeed, in the Pitchford and Snyder (2004) model, the progression of periods ended with some known probability (θ). It is unclear which situation obtains more frequently in natural settings. Clearly, some transactions are one-shot (e.g., the familiar example of tipping) while others are ongoing (e.g., Japanese Keiretsu). We hazard that there are many different kinds of interactions, and believe that there is value in exploring the space of alternatives. This should be a subject of future inquiry, a point to which we now turn.

9 Institutional design and future directions

The implications of these results for institutional design are somewhat mixed. The heterogeneity in trustworthiness among some people in the final iterations in the Increase Only condition is in particular worth close scrutiny. The mechanism seems to function reasonably well until the end game is reached. This suggests that institutional incentives to be trustworthy on the final round might aid the somewhat naturally occurring gradually increasing trust observed in the Unrestricted condition. This is interesting because it implies that with relatively minimal incentives—only the last round of an iterated transaction, as opposed to each round—the natural tendency to gradually increase trust could be less risky for first movers (Yamagishi 2003). One can imagine a slight change in the current design such that Player 2's who return less than Player 1's send on the last move of any iteration are fined. Such a one-time fine might prevent “unraveling” because Player 1's know that final moves entail relatively little risk.

Of course, adding such an element to the mechanism would translate into potentially large costs in the real world—the detection of such violations, the imposition of fines, and so forth. However, given the heterogeneity in behavioral strategies, it

might be that last-round defection in iterated transactions are sufficiently rare that enforcement costs would be minimal. We cannot directly address this issue, but we look forward to future work that, perhaps, explores mechanisms that focus on only final rounds of transactions to enable the gradual building of trust.

With respect to the heterogeneity observed, contrary to *ex ante* predictions, we observed important differences in our university subject pools. This should, of course, be of concern to researchers drawing from convenience samples at universities with the hope of generalizing to broader populations. Our data suggest caution in generalizing even from one university to another. In this sense, the growing research attention on varied populations in experimental bargaining games takes on additional importance (Buchan and Croson 2004; Carpenter et al. 2005; Henrich et al. 2005; Bohnet et al. 2005; Holm and Danielson 2005).

While the literature in economics is replete with examples of incentive compatible contracts to solve a variety of problems with trust and trustworthiness (see, e.g., Laffont and Tirole 1988; Moore and Repullo 1988), more work can be done to explore mechanisms that are better at dealing with endgame effects. Our results indicate that these can be, but not always are, potentially pernicious.

Finally, it is worth noting that new institutions designed around assumptions of rationality and financial self-interest are not always either desirable or necessary. The psychology of personal exchange seems to support trust and trustworthiness in a variety of settings, especially when trust can be built over a substantial time horizon. The key element of good institutional design surrounds understanding the relevant psychology that will mesh with institutions to elicit the desired social welfare-improving behavior. Future work should be aimed at clarifying these behaviors and developing new institutions, particularly those that are robust to the important population-level heterogeneity that is becoming increasingly apparent.

Acknowledgements This work has been supported by a University Research Foundation grant from the University of Pennsylvania to Kurzban and a grant from the International Foundation for Research in Experimental Economics to Wilson. We thank Tim Cason and two anonymous reviewers for comments and suggestions that have improved the paper. We would also like to thank Jeffrey Kirchner for valuable programming. All errors remain our own.

References

- Andreoni, J., & Samuelson, L. (2006). Building rational cooperation. *Journal of Economic Theory*, 127, 117–154.
- Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1), 122–142.
- Blonski, M., & Probst, D. (2004). *The emergence of trust* (Working Paper). University of Mannheim.
- Bohnet, I., & Zeckhauser, R. (2004). Trust, risk, and betrayal. *Journal of Economic Behavior and Organization*, 55(4), 467–484.
- Bohnet, I., Hermann, B., & Zeckhauser, R. (2005). *The elasticity of trust: Evidence from Kuwait, Oman, Switzerland, The United Arab Emirate, and the United States* (Working Paper). Harvard University.
- Bolton, G., Katok, E., & Ockenfels, A. (2003). *How effective are electronic reputation mechanisms? An experimental investigation* (Working Paper). University of Cologne.
- Buchan, N., & Croson, R. (2004). The boundaries of trust: Own and others' actions in the US and China. *Journal of Economic Behavior and Organization*, 55(4), 485–504.

- Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. New York: Sage.
- Carpenter, J., Burks, S., & Verhoogen, E. (2005). Comparing students to workers: The effects of social framing on behavior in distribution games. In J. Carpenter, J. List & G. Harrison (Eds.), *Research in experimental economics. Field experiments in economics* (pp. 261–290).
- Choi, S., Gale, D., & Kariv, S. (2006). *Sequential equilibrium in monotone games: Theory-based analysis of experimental data* (Working Paper). New York University.
- Croson, R., & Buchan, N. (1999). Gender and culture: International experimental evidence from trust games. *American Economic Review, Papers and Proceedings*, 89, 386–391.
- Dorsey, R. E. (1992). The voluntary contributions mechanism with real time revisions. *Public Choice*, 73, 261–282.
- Duffy, J., Ochs, J., & Vesterlund, L. (2006). *Giving little by little: Dynamic voluntary contribution games* (Working Paper 232). University of Pittsburgh.
- Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Journal of Economic Behavior and Organization*, 47(2), 268–298.
- Eckel, C., & Wilson, R. (2003). The human face of game theory: Trust and reciprocity in sequential games. In E. Ostrom & J. Walker (Eds.), *Trust and reciprocity: Interdisciplinary lessons from experimental research* (pp. 245–274). New York: Russell Sage Foundation.
- Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54, 293–315.
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817–868.
- Frank, R., Gilovich, T., & Regan, D. (1993). The evolution of one-shot cooperation. *Ethology and Sociobiology*, 14, 247–256.
- Friedman, J. W., & Hammerstein, P. (1991). To trade or not to trade; That is the question. In R. Selten (Ed.), *Game equilibrium models. I. Evolution and game dynamics* (pp. 257–275). Berlin: Springer.
- Gale, D. (1995). Dynamic coordination games. *Economic Theory*, 5, 1–18.
- Gale, D. (2001). Monotone games with positive spillovers. *Games and Economic Behavior*, 37, 295–320.
- Henrich, J., Boyd, R., Bowles, S., Gintis, H., Fehr, E., Camerer, C., McElreath, R., Gurven, M., Hill, K., Barr, A., Ensminger, J., Tracer, D., Marlow, F., Patton, J., Alvard, M., Gil-White, F., & Henrich, N. (2005). Economic Man' in cross-cultural perspective: Ethnography and experiments from 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795–815.
- Holm, H., & Danielson, A. (2005). Tropic versus Nordic trust: Experimental evidence from Tanzania and Sweden. *The Economic Journal*, 115, 505–532.
- Ishii, K., & Kurzban, R. (2007, in press). Public goods games in Japan: Cultural and individual differences in reciprocity. *Human Nature: An Interdisciplinary Biosocial Perspective*.
- Kurzban, R., & Houser, D. (2005). Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Science*, 102(5), 1803–1807.
- Kurzban, R., McCabe, K., Smith, V., & Wilson, B. J. (2001). Incremental commitment and reciprocity in a real-time public goods game. *Personality and Social Psychology Bulletin*, 27(12), 1662–1673.
- Laffont, J. J., & Tirole, J. (1988). The dynamics of incentive contracts. *Econometrica*, 56(5), 1153–1175.
- Lewicki, R. J., & Bunker, B. B. (1996). Developing and maintaining trust in work relationships. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in organizations: Frontiers in theory and research* (pp. 114–139). Thousand Oaks: Sage.
- Lindskold, S. (1978). Trust development, the GRIT proposal and the effects of conciliatory acts on conflict and cooperation. *Psychological Bulletin*, 85, 772–793.
- Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2), 224–253.
- Marx, L., & Matthews, S. (2000). Dynamic voluntary contribution to a public project. *Review of Economic Studies*, 67, 327–358.
- McCabe, K., Rigdon, M., & Smith, V. (2002). Cooperation in single play, two-person extensive form games between anonymously matched decision makers. In R. Zwick & A. Rapoport (Eds.), *Experimental Business Research* (pp. 49–67). Boston: Kluwer Academic.
- McCabe, K., Rigdon, M., & Smith, V. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, 52(2), 267–275.
- Moore, J., & Repullo, R. (1988). Subgame perfect implementation. *Econometrica*, 56(5), 1191–1220.
- North, D. C. (1981). *Structure and change in economic history*. New York: W. W. Norton and Company.
- North, D. C. (2005). *Understanding the process of economic change*. Princeton: Princeton University Press.

- Ortmann, A., Fitzgerald, J., & Boeing, C. (2000). Trust, reciprocity, and social history: A re-examination. *Experimental Economics*, 3, 81–100.
- Osgood, C. E. (1962). *An alternative to war or surrender*. Urbana: University of Illinois Press.
- Pillutla, M. M., Malhotra, D., & Murnighan, K. (2003). Attributions of trust and the calculus of reciprocity. *Journal of Experimental Social Psychology*, 39, 448–455.
- Pitchford, R., & Snyder, C. M. (2004). A solution to the hold-up problem involving gradual investment. *Journal of Economic Theory*, 14, 88–103.
- Rapoport, A., Stein, W., Parco, J., & Nicholas, T. (2003). Equilibrium play and adaptive learning in a three-person centipede game. *Games and Economic Behavior*, 43, 239–265.
- Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, 49, 95–112.
- Roberts, G., & Renwick, J. S. (2003). The development of cooperative relationships: An experiment. *Proceedings of the Royal Society of London B*, 270, 2279–2283.
- Roberts, G., & Sheratt, T. N. (1998). Development of cooperative relationships through increasing investment. *Nature*, 394, 175–179.
- Schelling, T. (1960). *The strategy of conflict*. Cambridge: Harvard University Press.
- Slonim, R., Engle-Warnick, J., & Helper, S. (2001). *Context in repeated trust games* (Working Paper). Case Western Reserve.
- Smith, V. (1991). *Papers in experimental economics*. Cambridge: Cambridge University Press.
- Triandis, H. C. (1995). *Individualism and collectivism*. Boulder: Westview Press.
- Watson, J. (2002). Starting small and commitment. *Games and Economic Behavior*, 38(1), 176–199.
- Weber, J. M., Malhotra, D., & Murnighan, J. K. (2004). Normal acts of irrational trust: Motivated attributions and the trust development process. In R. M. Kramer & B. Staw (Eds.), *Research in organizational behavior* (pp. 75–101). London: JAI Press.
- Yamagishi, T. (2003). Cross-societal experimentation on trust: A comparison of the United States and Japan. In E. Ostrom & J. Walker (Eds.), *Trust and reciprocity: Interdisciplinary lessons from experimental research* (pp. 352–370). New York: Russell Sage Foundation.
- Zak, P. J., & Knack, S. (2001). Trust and growth. *Economic Journal*, 111(470), 295–321.