A Matter of Trust: From Social Preferences to the Strategic Adherence to Social Norms

Joachim I. Krueger,[1] Adam L. Massey,[2] and Theresa E. DiDonato[1]

1 Department of Psychology, Brown University, Providence, RI, U.S.A.
2 Department of Mathematics, University of California, Los Angeles, CA, U.S.A.

Keywords
trust, reciprocity, social preferences, social norms, projection, reputation, self-image.

Correspondence
Joachim I. Krueger, Department of Psychology, Brown University, Box 1853, 89 Waterman St., Providence, RI 02912, U.S.A.; e-mail: joachim_krueger@brown.edu.

Abstract

In a mathematical analysis of the trust game, we show that utility-maximizing trustees should establish equal payoffs or return nothing depending on the strength of their social preferences (benevolence and inequality aversion). Trustors may invest any amount depending on their social preferences and their expectations regarding the trustees' preferences. For both types of player, empirical distributions of transfers are rather flat, however, and players' morality, but not their rationality, is judged in proportion to the money transferred. This pattern of findings suggests that people are primarily motivated by self-interest, and that they adhere to relevant social norms inasmuch as they can enhance their self-image or reputation as a moral person.

You have come into a little money, and you have the option of giving some of it to another person. Whatever you give, the invisible hand of the market—or the visible hand of the experimenter—will multiply, say by a factor of 3. The other person then decides how much to return to you. This arrangement is known as the investment or trust game (Berg, Dickhaut, & McCabe, 1995). The purpose of this game is to model human exchanges that are not enforced by contracts and that offer people no opportunities to reward or punish each other outside the context of the game itself. The relevance of trust and reciprocity has been noted in the contexts of economic behavior (Arrow, 1974), negotiation (Cialdini, 2001), organizational life (Kramer, 2007), and in the context of human exchanges in general (Homans, 1958).[1] Like other experimental

We thank Stephen Garcia, Rob Kurzban, and Judith Schrier for their helpful comments on a draft version of this manuscript.

[1]Trust is not a social nicety added onto otherwise contractualized exchanges. Trust-based exchanges have a longer evolutionary history than do contract-based exchanges. Yet, they are fragile. Explicit sanctioning mechanisms erode trust and trustworthiness (Mulder, van Dijk, De Cremer, & Wilke, 2006).

games, the trust game presents the decision dilemma most starkly when the players are anonymous and when the exchange occurs only once, that is, when decisions regarding investments and returns are not influenced by the players' history, their knowledge of each other, or their hopes and fears for the future. In other words, the trust game provides an experimental test of "blind trust."

The game-theoretic solution is that a rational self-interested trustee keeps all the money, and that a rational trustor, who knows this, invests nothing. The result is a deficient Nash equilibrium. The amount of money taken home is not what it could be. Empirically, the situation is less bleak because many trustors invest large sums, and many trustees return the invested amount. These trustees are conditional reciprocators, who reward the trustors at a cost to themselves (King-Casas et al., 2005). Many trustors seem to expect conditional reciprocity, or else they would invest nothing. The challenge to theory is to explain why trustors invest at all, and why trustees reciprocate.

In the search for answers, clues can be found in the writings of the Scottish philosophers who created the image of *homo economicus.* These philosophers also considered the role of moral sentiments. Adam Smith wrote one book on the virtues of rigorous self-interest (Smith, 1776/1869), and another one on the importance of moral passions such as sympathy, benevolence, and outrage (Smith, 1759/1976). Like Hume (1739/1978), Smith believed that humans care about the well-being of others. "How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render his happiness necessary to him, though he derives nothing from it except the pleasure of seeing it" (Smith, 1759/1976, p. 9). The notion of moral sentiments has stimulated the development of social preference theories that seek to go beyond self-interest in explaining interpersonal behavior (e.g., Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999; Rabin, 1993; Van Lange, 1999). These theories assume that decisions not only depend on a person's own monetary payoffs, but also on the payoffs reaped by others (benevolence), and on the differences between own and others' payoffs (inequality aversion).

Social preference theories encourage the study of individual differences. Messick and McClintock (1968) distinguished among several types. Individualists are motivated only by self-interest: they care about maximizing their own payoffs. Altruists are motivated by benevolence: they care about maximizing the payoffs of others. Competitors are negatively motivated by inequality aversion: they seek to maximize the difference between their own and the other's payoffs even if that means that they have to forego a higher possible payoff for themselves. Finally, cooperators are motivated by all three types of preference: they are self-interested, but they also care about others' payoffs and about fairness. Cooperators are of special interest in the trust game because, on the face of it, they are they only type that, based on the mixture of their preferences, might invest or reciprocate partially. Building on interdependence theory (Kelley & Thibaut, 1978), Van Lange (1999) proposed that social preferences can be modeled by the weights people assign to self-interest, benevolence, and inequality-aversion. When all three weights are estimated, the given payoffs in an experimental game can be transformed into utilities. "Choices are then made from the transformed matrix that maximizes expected utility" (Van Lange & Liebrand, 1991, p. 275). Van Lange (1999) has suggested that noncooper-

ative games such as the prisoner's dilemma can become games in which the cooperative option dominates (but see Krueger, 2007).

The trust game can be construed as a sequential prisoner's dilemma. As in the prisoner's dilemma, a cooperative player (trustor) accepts the risk of being exploited. Unlike in the prisoner's dilemma, however, the trustee can knowingly exploit or reciprocate. The question is whether social preferences can transform given payoffs into utilities that make the transfer of money a rational strategy. We begin our analysis with the trustees because their decisions are what the trustors need to anticipate before deciding how much to invest. The following variables are at play:

E_1: The trustor's endowment.
E_2: The trustee's endowment.
p: The multiplier of E_1 that determines the proportion invested ($0 \le p \le 1$).
m: The multiplier of pE_1 that determines how much the trustee receives.
q: The multiplier of $E_2 + mpE_1$ that determines how much the trustee returns ($0 \le q \le 1$).
w_1: Benevolence, i.e., the weight given to the trustor's payoff ($0 \le w_1 \le 1$).
w_2: Inequality aversion, i.e., the weight given to the difference between the trustee's and the trustor's payoff ($0 \le w_2 \le 1$).[2]

The trustee's payoff is $E_2 + mpE_1 - q(E_2 + mpE_1)$, and the trustor's payoff is $q(E_2 + mpE_1) + E_1 - pE_1$. After transformation, the trustee's utility is his or her own payoff plus the trustor's weighted payoff minus the weighted absolute difference between the two payoffs. Thus, the trustee seeks to maximize

$$F(q) = E_2 + mpE_1 - q(E_2 + mpE_1) + w_1[q(E_2 + mpE_1) + E_1 - pE_1] - w_2|E_2 + mpE_1 - 2q(E_2 + mpE_1) - E_1 + pE_1|.$$

For simplicity, we begin by assuming that E_1 = \$10, E_2 = \$0, m = 3, and $w_1 = w_2$. It is evident that the trustee will keep all the money if $w = 0$. Conversely, if $w = 1$, and if the trustor invests all, the trustee will return $\frac{mpE_1}{2}$. Now consider $w = 0.5$ as an empirically plausible level of benevolence and inequality aversion. If the trustor invests all and the trustee returns half, the trustee's utility is \$22.5, namely \$15 + \$7.5 − \$0. This is the maximum utility, which can be seen by considering departures from equality. A trustee who returns only \$14 realizes a utility of \$16 + \$7 − \$1 = \$22. Compared with the even-split scenario, the trustee's own payoff is increased by \$1, the weighted payoff of the trustor is reduced by \$0.5, and the weighted inequality is reduced by \$1. The trustee would be indifferent if the sum of the two weighted reductions were equal to the increase in own payoff. This would happen if $w = \frac{1}{3}$. For equal redistribution, the trustee's utility would be $\$15 + \frac{\$15}{3} - \$0 = \$20$; for unequal redistribution, it would be $\$16 + \frac{\$14}{3} - \frac{\$2}{3} = \$20$.

[2]Whereas Van Lange (1999) allowed the weight given to own payoffs (i.e., self-interest) to vary, we set this term to 1 in order to express the strength of benevolence (w_1) and the strength of inequality aversion (w_2) in proportion to self-interest. This simplification was empirically justified by the finding that cooperators are no less self-interested than individualists or competitors.

This illustration leads to a surprising conclusion: There exists a critical level of social preference below which the trustees will not reciprocate. Above the critical point, they will reciprocate the amount needed to eliminate inequality. This conclusion is at odds with the wide empirical distribution of trustees' transfers, suggesting the social preference models provide a poor—or at best incomplete—account of reciprocating behavior.

Overview

The present article is organized in three parts. In part 1, we mathematically derive the trustee's indifference point, and prove that below this point reciprocation should not occur, whereas above it, equality should be sought. It follows that self-interested trustors should invest all or nothing, depending on how strong they believe the trustees' preferences to be. Arguably, trustors may also be motivated by social preferences. Some participants transfer money even when they know that their partners are not allowed to reciprocate (Cox, 2004; Cox & Deck, 2005). We therefore derive the investment that maximizes the trustor's utility. To anticipate the key finding, partial investments can maximize the trustor's utility for certain combinations of social preferences and expectations of reciprocity.

In part 2, we develop the idea that players in the trust game are, in part, motivated by the social images they project to themselves and others. Building on evidence that the dimensions of morality and competence are fundamental to social perception (Fiske, Cuddy, & Glick, 2006; Wojciszke, 2005), we present a study in which participants judge the personalities of trustors and trustees who transfer varying proportions of the available funds. The goal of this study is to see whether personality impressions capture the regularities predicted by game theory or the social preference model. We expected that perceptions of morality would increase with the dollar amount transferred. With regard to perceptions of competence, we consider four hypotheses. Of these, the hypothesis of greatest interest follows from our quantitative analysis. Namely, a trustee who returns either nothing or who returns the amount that establishes equality can be seen as rational. A partial reciprocator cannot be seen as rational because partial returns cannot maximize utility. Conversely, a trustor can be seen as rational regardless of the invested amount. The other three hypotheses are elaborated when the study is described.

In part 3, we review the implications of our mathematical modeling and our person-perception study for the social preference model. Noting the insufficiency of the model, we review recent research that, as a whole, supports the view that people adhere to social norms strategically. Norms of reciprocity (and to a lesser extent, norms of trust) need to be activated by contextual cues. Both, selective norm adherence and strategic norm suspension, can serve to sustain positive self-images and social reputations. Social exchanges are thus more fragile than dispositional theories such as the social preference model would imply. We sketch a more integrative model that takes strategic norm adherence into account, and we note its limitations as a prescriptive model.

The Trustee

The trustee faces the question of what proportional transfer q maximizes utility. We consider separately inequalities implying guilt (i.e., when the term inside the absolute value >0) and inequalities implying envy (i.e., when the term inside the absolute value <0; Fehr & Schmidt, 1999). Equality is the boundary between these two regions. In the expression $E_2 + mpE_1 - E_1 + pE_1 = 2q^* (E_2 + mpE_1)$, q^* is the proportion of the trustee's funds that yields equality. Solving for q^*, we find that $0.5 - \frac{E_1}{2(E_2 + mpE_1)} + \frac{pE_1}{2(E_2 + mpE_1)} = q^*$. Note that differentiation is not possible when $q = q^*$ because the absolute function is not differentiable at zero.

First, consider the case in which the term in the absolute value is positive. For the range of $0 < q < q^*$, the trustee's utility can be written as $(1 - w_2)(E_2 + mpE_1) + q(w_1 + 2w_2 + mpE_1) + q(w_1 + 2w_2 - 1)(E_2 + mpE_1) + (w_1 + w_2)(E_1 - pE_1)$, and the derivative of this expression is $\frac{dF}{dq} = (w_1 + 2w_2 - 1)(E_2 + mpE_1)$. If $w_1 + 2w_2 = 1$, the derivative is zero, and the trustee is indifferent about the size of the return. If, however, $w_1 + 2w_2 < 1$, the derivative is negative, which means that utility is maximized at $q = 0$; if $w_1 + 2w_2 > 1$, the derivative is positive, with utility maximized at $q = q^*$.

Second, consider the case in which the term in the absolute value is negative. For the range of $q^* < q \leq 1$, the trustee's utility can be written as $(1 + w_2)(E_2 + mpE_1) + q(w_1 - 2w_2 - 1)(E_2 + mpE_1) + (w_1 - w_2)(E_1 - pE_1)$, and the derivative of this expression is $\frac{dF}{dq} = (w_1 - 2w_2 - 1)(E_2 + mpE_1)$. This derivative cannot be positive because $w_1 - 2w_2$ cannot be >1. Choosing $q > q^*$ only decreases utility.

These derivations confirm what the introductory example suggested. Depending on the total strength of their social preferences, trustees can maximize their utilities by returning nothing or by establishing equality. This result is at odds with the findings of the typical trust game, where the distribution of returns is rather flat (Berg et al., 1995; Burks, Carpenter, & Verhoogen, 2003). The derivations also highlight the greater impact of inequality aversion. If w_2 were zero, even perfect benevolence ($w_1 = 1$) could not induce the trustee to return funds. If, however, w_1 were zero, any w_2 above 0.5 would have that desired effect. Finally, the model implies that benevolence and inequality-aversion are confounded when the latter takes the form of guilt, and work against each other when the latter takes the form of envy. Since most people are less averse to guilt than to envy (Adams, 1963), it follows that benevolence and inequality aversion are positively correlated (Van Lange, 1999).

The Trustor

A self-interested trustor who knows—however intuitively—the constraints on the returns imposed by the trustee's social preferences either invests nothing or the full amount. The challenge to the trustor is to correctly predict whether the trustee's preferences are strong enough to seek equality. Like trustees, however, trustors may differ in the weights they place on benevolence and inequality aversion. Hence, trustors seek to maximize their own payoffs plus the weighted trustee's payoff minus the weighted difference between the two payoffs or

$$G(p) = q(E_2 + mpE_1) + E_1 - pE_1 + w_1[E_2 + mpE_1 - q(E_2 + mpE_1)] - w_2[E_2 + mpE_1 - 2q(E_2 + mpE_1) - E_1 + pE_1].$$

The trustor's task is complicated by the fact that the strength of reciprocity, q, is unkown. We begin by deriving the value of p that eliminates the payoff difference. Simple algebra shows that the trustor anticipates payoff equality if $\tilde{p} = \frac{2qE_2 + E_1 - E_2}{mE_1 - 2qmE_1 + E_1}$. The psychological implication of this result is that $\tilde{p}$ is correlated with q. The more reciprocity a trustor expects, the larger is the investment that yields expected equality.

Again, we consider two cases. In the first case, the term in the absolute value is positive, which implies envy from the trustor's perspective. For the range of $\tilde{p} \leq p \leq 1$, the derivative of $G(p)$ is $\frac{dG}{dp} = E_1(mq + w_1m - w_2m - 1 + 2qmw_2 - qmw_1)$. The sign of the derivative depends on the strength of the trustor's social preferences. If $w_1(m - qm) + w_2(2qm - m) < 1 - mq$, the derivative is negative and the smallest value for p (i.e., $\tilde{p}$) maximizes the trustor's utility. If, however, the inequality is reversed, the derivative is positive, and $p = 1$ maximizes utility. Finally, if there is no inequality, the derivative is 0 and the trustor indifferent. Across a range of plausible values, stronger social preferences and stronger expectations of reciprocity increase the likelihood of $p = 1$ being the utility maximizing investment. Indeed, this is the result obtained for most plausible values. Investment of size $\tilde{p}$ can be best, but only if social preferences are very weak and if expectations of reciprocity are very low.

In the other case, the term in the absolute value is negative, which implies trustor's guilt. For the range of $0 \leq p \leq \tilde{p}$, the derivative of $G(p)$ is $\frac{dG}{dp} = E_1(mq + w_1m - w_1qm + w_2m - 2w_2qm + w_2 - 1)$. Again, the sign of the derivative matters. If $mq + w_1m - w_1qm + w_2m - 2w_2qm + w_2 - 1 < 0$, the derivative is negative, and utility is maximized for $p = 0$. If the same expression >0, the derivative is positive, in which case $\tilde{p}$ maximizes utility. In this case too, stronger social preferences increasingly yield larger instead of smaller investments ($\tilde{p}$ rather than 0) as the maximizing investment. Null investments are again restricted to the combination of very weak preferences and very weak expectations of reciprocity.

The derivations show that for trustors, partial investments can be in line with the social preference model. However, the trustors' inferential task is highly complex. Not only do they need to know their own preferences, they must also generate expectations regarding the trustees' intention to reciprocate. If the trustors assume that the trustees are rational utility maximizers, they can only assume that the rate of reciprocity will be 0 or q^*. If they do not assume rationality, they also need to consider the possibility of reciprocity between 0 and q^*. However heuristically trustors solve this task, they empirically tend to end up making smaller investments than the social preference model suggests.

Experiment: Perceptions of Trustors and Trustees

The premise of this experiment is that the decisions trustors and trustees make will affect the impressions others have of them. Whereas game theory is only concerned with the modeling of exchange behavior, social preference models also address the question of how people react to behavior that violates norms or expectations (Gintis, Bowles, Boyd, & Fehr,

2005). We assess how trustors and trustees are perceived in terms of their morality and competence (i.e., rationality). Specifically, we first ask how strongly perceptions of morality are associated with the size of the transfer, and then ask how well perceptions of competence map onto the predictions of game theory or the social preference model.

Hypotheses

One central facet of moral behavior is that it benefits others (Peeters, 2002). Therefore, a close association between judgments of players' morality and the size of their transfers is to be expected (Singer et al., 2006). The social preference model further suggests that the player's role makes a difference. A trustee's decision to reciprocate can be seen as a direct reflection of social preferences, and hence be judged in moral terms. In contrast, a trustor's decision to invest is ambiguous because this player stands to gain only if the trustee reciprocates. The trustor can bet on the trustee having the kinds of social preferences that the trustor him- or herself does not need to have. Therefore, social preference theories predict that judgments of the trustee's morality will increase more steeply with the amount transferred than will judgments of the trustor's morality.

The social preference model casts people as utility-maximizers. The model implies that people who fail to act in accordance with their preferences should not be considered rational. Our mathematical derivations specifically show that a trustee's partial reciprocation cannot be mapped onto any combination of preferences. One version of the social preference model aligns rationality with morality. On this view, larger returns signify greater "collective rationality" (Van Lange & Liebrand, 1991). Inasmuch as naïve participants share this view, they may use their perceptions of morality as a cue for making judgments of competence. Both of these hypotheses contrast with the game-theoretic view, which regards reciprocity as a violation of self-interest. Therefore, judgments of competence should decrease with the amount returned. If so, the resulting pattern would resemble the so-called "compensation effect," according to which perceptions of morality and competence are negatively related (Judd, James-Hawkins, Yzerbyt, & Kashima, 2005). A final possibility is that returns are not relevant for judgments of competence at all. Conceivably, people construe the trust game exclusively in moral terms, in which case judgments of competence would not vary with the amount transferred. For the trustor, the same four hypotheses are tested. The first hypothesis, which assumes low ratings for partial investments, is not likely a priori, however, because the trustor can justify partial transfers.

Social Projection

Another objective of this study is to elicit participants' predictions of what they themselves would do as a trustor or as a trustee who receives a full investment. The significance of these measures is twofold. First, participants act as observers when judging a trustor or a trustee. Yet, they may spontaneously consider what they would do if they were in either one of these roles, and their judgments of others might be moderated by their egocentric standards of what they believe to be the "correct" response. Second,

investments and returns tend to be positively correlated across pairs of players (Cox, 2003). We therefore expect a positive correlation when transfers are generated by the same people.

The trustor's problem is to predict the trustee's willingness to reciprocate. Some trustors may have formed beliefs about the distribution of rates of reciprocity from their experience with social exchanges. Others are less knowledgeable or less willing to transfer the lessons of real-world exchanges to a novel laboratory situation. Still, these players can read the trustees' minds by reading their own. They can ask themselves how much they would return, given an investment of a certain size, if they were the trustees. By projecting their own intended transfers onto others, the trustors can capitalize on the fundamental similarities across humans (Humphrey, 1976). It is well known that players in noncooperative games project their own choices onto others. In the prisoner's dilemma, most cooperators expect cooperation, and most defectors expect defection (Dawes, McTavish, & Shaklee, 1977). Likewise, general attitudes toward social exchange are projected such that, compared with individualists, cooperators are more likely to expect others to cooperate (Kuhlman & Wimberley, 1976). Yet, the predictions are post-choice projections. The person's own strategy is assumed to be a stable disposition, which is expressed in the game and then projected onto others.

Projection can also occur at the prechoice stage (Krueger & Acevedo, 2005). The trustor's own sense of how much they think they would return if they were the trustee may affect their expectations about the trustees they face. Because the expected value of the game increases with expectations of reciprocity, people may opt to invest to the degree that they project to others. By using this heuristic, they come to enact social exchanges that are, on average, more efficient than they would be if they thought the behavior of others were unpredictable. Yet, there is reason to believe that the heuristic of social projection is often not fully exploited. For example, if players thought that trustees will return 80% of the maximum amount, because that is what they themselves would do in that role, they would do best if they invested everything. If they match their investment rate with the expected rate of reciprocity, they do worse.[3]

Methods

Ninety-four female and 44 male undergraduate students at Brown University participated in a class setting. Their mean age was 19.5 years with a range from 18 to 24 years. Each received a sheet of paper with the following instructions:

> In the Trust Game, person A receives \$10. S/he can turn any amount (\$0 to \$10) over to person B. The experimenter triples this amount. Person B then decides what portion of this

[3]Suppose $E_1 = 10$, $E_2 = 0$, $p = 0.8$ or 1, $m = 3$, and $w = 0$. The rate of reciprocity, q, is the maximum (i.e., equality yielding) amount divided by the money held by the trustee, or $\frac{\left[\frac{E_1 - pE_1 + mpE_1}{2}\right] - E_1 - pE_1}{mpE_1}$. In the present example, $q = 0.483$ for $p = 0.8$ and $q = 0.5$ for $p = 1$. If the trustee returns 80% of the available assets, the trustee ends up with \$10.08 or \$12, respectively, for $p = 0.8$ and $p = 1$.

total to return to person A. Assume that both A and B act anonymously. They do not know each other and will not interact in the future.

How would you rate the personality of a trustor (person A) and the personality of a trustee (person B) given their choices. Rate one trustor and one trustee described below on six trait-descriptive adjectives. Note that these two individuals are not necessarily partners in the same game. Use a 9-point scale (1 = the trait does not apply at all; 9 = the trait applies very well). Circle the number that represents your judgment.[4]

Next, a trustor who invested \$0, \$5, or \$10 was described, followed by a trustee who had received a full investment, and who returned \$0, \$10, or \$15. Each possible combination of investment and return was used with about the same frequency. Participants then rated the trustor and the trustee on six trait-descriptive adjectives, namely Generous, Intelligent, Moral, Naïve, Rational, and Selfish. These adjectives were culled from scales developed to assess impression formation in the context of experimental games (Krueger & Acevedo, 2007).

After rating the two players, participants were invited to imagine being in the trust game themselves. They indicated the dollar amount they would invest as a trustor. Then, assuming they were playing with a trustor who had invested all (i.e., \$10), they indicated how much they would return of their available funds (i.e., \$30).

Results and Discussion

We first examined the correlations among the rating variables to see if they adequately fell into the clusters of morality and competence. The pattern of correlations was similar for the two players, and the correlation across the 15 correlations was .94. Whereas ratings on the three morality traits were highly intercorrelated ($M = .65$ after r-Z-r transformation), ratings on the three competence traits were not, $M = .27$. The latter result occurred because naïvité was only weakly associated with intelligence and rationality. To construct scales of the same length, ratings on the adjectives naïve and moral were dropped, which resulted in reliability coefficients of .66 and .55, respectively, for morality and competence. The two scales were independent of each other, with $r(136) = .09$ and .12, respectively, for ratings of the trustor and ratings of the trustee (both $p > .15$).

Hypothesis tests were performed on the average ratings on each dimension. Although each participant rated one trustor and one trustee, we treated these judgments as independent because the portrayed investments and returns were manipulated independently. Because of the large sample size, there was little loss of statistical power. Hence, the data-analytical model was a 2 (Role: trustor vs. trustee) by 3 (Transfer: low, medium, large) factorial ANOVA. Figure 1 shows the relevant means and the standard errors.

[4]A show up fee for the trustee was not mentioned because (a) it did not affect the analytical properties of the social preference model, and because (b) empirical rates of reciprocation suggest that trustees are only concerned with dividing the multiplied investment, but not their total wealth.

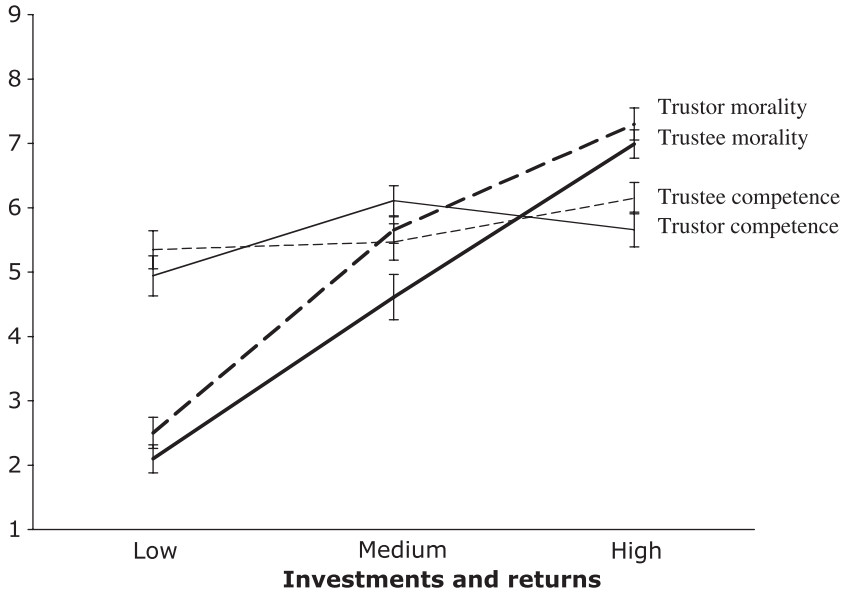

Figure 1. Average ratings of trustor and trustee on morality and competence as a function of amount offered.

Morality

As predicted, perceptions of morality varied strongly with the dollar amount transferred, $F(2, 271) = 185.24$, $\eta^2 = 0.58$. Players who transferred a partial amount were seen as more moral than players who transferred nothing, $F(1, 178) = 123.47$, and as less moral than players who transferred a lot, $F(1, 186) = 64.44$. Moreover, there was a statistical effect of the player's role. Trustors were seen as more moral than trustees, $F(1, 271) = 8.04$, $p = .005$, $\eta^2 = .03$. Although this effect was small, its direction ran counter to our original expectation that trustees would be seen as more moral, especially if they reciprocated strongly.

Three possible explanations for this result suggest themselves. First, trustors might have been seen as more moral because their investments were proportionally larger than the trustees' returns.[5] This hypothesis cannot explain why the size of the transfer did not moderate the differences between ratings of trustors and trustees, $F = 1.22$. A trustor who invests nothing should be judged to be as immoral as a trustee who returns nothing. A second idea is that only the trustees are under strong normative pressure (Gouldner, 1960). Whereas trustees are morally obligated to reciprocate, trustors have no moral obligation to invest (Bohnet & Zeckhauser, 2004). Inasmuch as observers discount norm-adherence from their judgments of moral credit, they might see the trustors as more moral. Again, this hypothesis would imply an interaction

[5]We thank Lucia Donatelli for suggesting this hypothesis.

between level of contribution and role. A third explanation draws attention to the different psychological situations faced by the two players. The trustor makes a decision under uncertainty, and bears a considerable risk (Luhmann, 1988; Malhotra, 2004). Being uncertain about the trustee's preferences, the trustor has to rely on probability estimates when trying to assess the expected value of the game. In contrast, the trustee makes a decision without risk. Knowing that his or her decision will conclude the game, the trustee only needs to consult accessible social preferences, and calculate how to maximize the utility of the game. If participants considered the presence of risk to be a unique burden for the trustor, they might give extra moral credit to anyone playing this role.[6]

Competence

The mean competence ratings in Figure 1 are consistent with the hypothesis that the size of the transfer does not affect perceptions of rationality or intelligence. The statistical significance of this variable, $F(2, 271) = 4.41$, $p = .013$, is overshadowed by the fact that the effect size was rather small, $\eta^2 = .03$. Simple comparisons showed only that players who transferred a partial amount were seen as more competent than players who transferred nothing, $F(1, 186) = 5.37$, $p = .02$, but they were not seen as less competent than players who transferred a lot, $F < 1$. The player's role had no direct effect, $F < 1$, but it interacted with the size of the transfer, $F(2, 271) = 3.06$, $p = .049$. Again, however, the effect was small, $\eta^2 = .02$, making its practical significance doubtful. It is noteworthy that the mean competence ratings clustered around the midpoint of the 9-point scale, which further supports the idea that participants did not construe the trust game as a context in which rationality is revealed. This conclusion is reinforced by the small size of the standard errors and their homogeneity across levels of contribution and across players' roles. Evidently, the mean judgments did not mask strong but opposite impressions of competence.

It is critical to point out what did *not* happen. Contrary to how game theory suggests that people ought to be judged, participants did not associate a player's competence negatively with the size of the transfer. Contrary to the implications of the social preference model, trustees who returned intermediate amounts were not rated as least competent. Finally, the evidence for the idea that competence ratings, like morality ratings, increase with the size of the transfers was so weak that it lacks practical significance.[7]

---

[6]Stephen Garcia suggested the related idea that participants are sensitive to the direction of social comparisons. The trustors face upward comparisons, whereas the trustees face downward comparisons. Inasmuch as they see the former as psychologically more stressful than the latter, participants may reward the trustors by attributing greater morality to them. Finally, it should be noted that the main effect of role is confounded with an order effect, as the trustee was always judged after the trustor. Because the trust game is played as a sequence of moves, we did not counterbalance the order of the ratings.

[7]Across participants and experimental conditions, judgments of morality and competence were fairly independent ($r = .08$ and .12, respectively, for trustors and trustees). It was therefore legitimate to perform separate ANOVAs on these two variables. Nonetheless, we performed regression analyses to ensure that these small correlations did not qualify the results. Judgments of morality were regressed on condition and on judgments of competence; judgments of competence were regressed on condition and judgments of morality. Without exception, the regression findings were the same as the ANOVA findings.

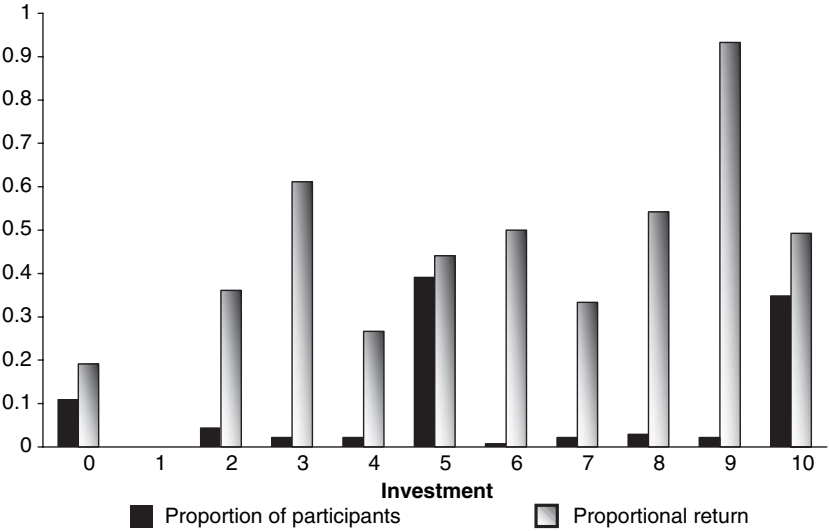

Figure 2. Proportion of participants choosing each level of investment and mean fraction of return relative to the maximum of $30.

Own Investments and Returns

The solid bars in Figure 2 show that the distribution of investment decisions was trimodal, with most participants choosing either to invest nothing, half, or the full amount. Few participants (16.67%) chose investments between these modes. The shaded bars show the average return decisions for each level of investment as a proportion of the maximum amount of $30. Ignoring the unbalanced frequency distribution, the correlation between investments and returns was .67. The correlation across all participants was $r(136) = .42$, $p = 3^{-7}$. As expected, investments and returns were closely related for individual players.

The final question was whether participants' own predicted transfers affected how they judged other players.[8] To find out, investments and returns were standardized, averaged, and standardized again. Conditions were coded as −1, 0, and 1, respectively, for low, medium, and high transfers. Then, the standardized judgments of morality and competence were regressed on condition, own transfers, and the cross-products of the two. For morality, these analyses replicated the ANOVA findings, and showed no moderating effect of own transfers ($\beta = -.07$ and $-.05$, respectively, for the trustor's and the trustee's judged morality). Likewise, the trustee's judged competence was not moderated by own transfers, $\beta = -.08$. There was only a statistical moderator effect for the trustor's

[8]In our study, participants' responses to the trust game were simulated rather than real. Camerer and Hogarth (1999) noted that real incentives improve performance and reduce the response variability, but that "the modal result has no effect" (p. 7). The validity of our data is supported by the finding that the distribution of simulated transfers resemble empirical distributions reported in the literature.

competence, $\beta = -.18$; $p = .02$. Inspection of the within-condition correlations between own transfers and judged competence revealed that this interaction was produced by one unforeseen finding. Whereas own transfers were unrelated to judged competence when the trustor returned nothing or a partial amount (both $r = -.01$), own transfers were negatively related to the trustor's judged competence, $r = -.47$. In other words, participants who themselves thought they would transfer less perceived a fully investing trustor as more competent. This finding does not readily fit with any of the a priori hypotheses. Because it emerged against the background of own transfers being irrelevant to this person perception task, we consider this finding an anomaly for the time being.

Discussion

The trust game provides a fertile testing ground for normative and descriptive theories of social exchange. As evidence for the limitations of traditional rational-choice models is mounting, many proposals have emerged to enrich theory with cognitive, behavioral, or emotional parameters (Camerer, 2003; Colman, 2003; Gintis et al., 2005). Social preference theories are attractive because they give form to the idea that many people care about the well-being of others and dislike inequality. Our quantitative analysis revealed the implications of one such model (Van Lange, 1999). Trustees must know their social preferences. When they do, they should return nothing or establish payoff equality. Self-interested trustors who know this should invest all or nothing depending on their assumptions about the trustees' social preferences. However, trustors' own social preferences may justify partial investments.

Some trustors may have enough information to formulate expectations regarding the trustee's social preferences. Others may lack such information, but heuristically project their own preferences onto the trustee. The trustors' maximum expected value of the game is then $\frac{mE_1 + E_2}{2}e$, where e designates the perceived probability that the trustee fully reciprocates. A self-interested trustor should therefore invest all if $e > \frac{2}{m}$. Here, the multiplier matters, suggesting that with larger prospective gains, trustors should invest even when the perceived probability of trustworthiness is low.

However rational a preferences-times-projection model might be, it does not fully account for all the empirical evidence. Many trustees reciprocate partially, and some trustors invest more than the formal model can justify. The theoretical challenge to explain these behaviors thus remains. Although social preference models go beyond pure self-interest models, they remain consequentialist. What matters is still the weighted outcome combined with the probability with which it is believed to occur. We now ask to what extent the incorporation of social norms can improve explanations of behavior in the trust game.

Strategic Norm Adherence

The norm of reciprocity readily applies to the trustees. On average, trustees return (nearly) as much as trustors invest. Therein, however, lies a problem. Matching the trustors' investments preserves an imbalance. The trustees get richer, while the trustors

barely break even. This ambiguity does not exist in a sequential prisoner's dilemma, where a player who matches cooperation with cooperation not only reciprocates but also achieves payoff equality. Another possibility is that some people have internalized deontological norms, which demand that prosocial acts be performed as a duty and without regard for utilitarian benefits (Kant, 1785/1964). Still, a strictly nonconsequentialist theory of norm-adherence is as implausible as is a purely utilitarian theory of social preferences. A nonconsequentialist theory also suggests that people either contribute all or nothing, depending on whether they have internalized the norm.

A third possibility, which we consider promising, is that partial transfers in the trust game reflect a compromise between utilitarian and norm-abiding tendencies (Murnighan, Oesch, & Pillutla, 2001). It is here that the data from our scenario study are directly relevant. A trustee might be interested in projecting the image of a norm-abiding person because an act of norm adherence can itself have a positive utility. Some of the early moral philosophers recognized that many people are pleased with themselves when they meet social demands because they anticipate and value the approval of others. By enabling benevolent behavior, the need for approbation (Adam Smith) or the love of fame (David Hume) take self-interest beyond the limits of narrow economic calculations. Investments and reciprocations in the trust game may be like other costly expressive acts (e.g., voting): they are nonutilitarian from a dollars-and-cents perspective, but not without hedonic benefits. The perception of players in moral rather than rational terms suggests the impact of social norms, and thus people's expectation that they will be judged in terms of norm-adherence.[9]

Once social norms of interpersonal behavior are widely accepted, they generate their own consequentialist implications. This leads to an ambiguity. People may adhere to norms because they have internalized them, or because they recognize the social benefits of adherence. Experimentally, this ambiguity can be removed. We now review four sets of studies that explore what happens when people have an opportunity to act selfishly, while preserving the image of being socially responsible.

The first piece of evidence comes from a study on the ultimatum game.[10] Falk, Fehr, and Fischbacher (2003) limited proposals either to a choice between a \$8/\$2 split and a \$5/\$5 split or a choice between a \$8/\$2 split and a \$2/\$8 split. Responders were less likely to reject a \$8/\$2 split in the latter case than in the former, presumably because they realized that when the responder cannot establish equality, selfishness is excusable. By analogy, suppose a trustor must choose between investing \$10 and \$8 or between \$8 and \$5. A trustor choosing \$8 will probably be seen as more moral in the latter case than in the former, and a trustee may be inclined to return more. Likewise, suppose a trustee has a choice between returning \$15 and \$10 or between returning \$10 and \$5.

[9]Perceptions of players in the prisoner's dilemma show a similar pattern. Cooperators are seen as far more moral than defectors, whereas defectors are seen as only marginally more competent than cooperators (Krueger & Acevedo, 2007).

[10]In the ultimatum game, the proposer makes an offer as to how to split the stakes (typically \$10). If the responder accepts, the stakes are divides accordingly. If the responder rejects the offer, neither player receives anything.

Again, if the choice of returning $10 occurred in the latter context, perceptions of morality would be enhanced.

Second, perceptions and anticipated perceptions depend on available knowledge. Kagel, Kim, and Moser (1996) disabled the assumption of common knowledge in the ultimatum game by telling proposers that, unbeknownst to the responders, any dollar they received would be multiplied by three. Most proposers still offered an even split, which naturally was accepted by most responders. Only the *appearance* of fairness mattered, which runs counter to the idea of stable social preferences. By analogy, one can expect that a trustee who knows that the experimenter will secretly add to his or her final payoff will be tempted not to share these extra funds with the trustor. Indeed, the experimental evidence shows that most trustees do not consider their show up fee, E_2, to be part of the wealth to be shared with the trustor. When they do establish equality, it is usually only the equality based on the trustor's endowment and transfer. This finding suggests that trustees engage in compartmentalized (and self-serving) mental accounting (Thaler, 1999).

Third, for norms to have an effect, they need to be activated. Cox (2003) found that reciprocal transfers in the trust game were larger than transfers in the dictator game.[11] This makes good sense because only in the trust game, but not in the dictator game, can the norm of reciprocity be invoked. In the trust game, it is critical that the trustees see investments as intentional acts (McCabe, Rigdon, & Smith, 2003). When investments appear to be involuntary, the trustees do not feel bound by the norm. The behavior of dictators becomes even more selfish when the residual normative demand to give is weakened. Dana, Weber, and Kuang (2007) asked players to choose between $6 and $5 for themselves, while there was a 50% chance the receiver would get $5, otherwise $1. The ingenious feature of this design was that the dictator could find out at no cost what the receiver would get, given his or her own choice. Whereas most players chose the $5/$5 split in the traditional, no-uncertainty game, most players claimed $6 in the modified game without asking the receiver's payoff to be revealed. Under the guise of uncertainty, they felt free to ignore the social norm of fairness, presumably because their self-imposed ignorance of the receiver's payoff allowed them to maintain a moral self-image. Dana, Cain, and Dawes (2006) added an "opting-out" condition to the dictator game. Many participants chose to privately receive $9 (without the opportunity to share), thereby avoiding the normative pressure to divide a stake of $10. By analogy, we can surmise that given the opportunity to preserve a moral self-image, many trustees will disregard social preferences or norms and resort to self-interest.

These examples suggest that people seek to protect their social reputations and moral self-images. When they have "an out" to pursue their self-interest without damaging their reputations, they take it. Studies varying the presence of absence of observers show most directly that being potentially judged by others is a powerful social inducement to act cooperatively. Haley and Fessler (2005) found that visual and

[11]In the dictator game, one player, the dictator, divides the stakes as he or she sees fit. The other player has no decision to make.

auditory cues of social monitoring increased generosity in a dictator game. Kurzban, DeScioli, and O'Brien (2007) found that third-party punishment of defectors at a cost to the self is strong only when the punishment was witnessed. The finding that some transfers occur even in the absence of observers suggests that some people have transformed their reputational concerns into a private motive to preserve a moral self-image. As Pillutla, Malhotra, and Murnighan (2003, p. 450) suggested, "people want to see themselves positively, even when their actions are anonymous." Such concerns and motives need not be irrational. People who act with perfect selfishness can realistically anticipate social rejection. Censure or ostracism are highly stressful experiences (Dickerson & Kemeny, 2004), and averting them is a potent negative reinforcement for cooperation.

The final set of evidence comes from studies showing the contextual malleability of economic exchanges. Even arbitrary differences in the players' label are sufficient to affect trust and generosity. "Partners" receive more cooperation than "opponents" (Messick & McClintock, 1968). Such labels signal differences in social distance. Over a range of social distance (i.e., from close friend or relative to mere acquaintance), transfers decrease hyperbolically (Jones & Rachlin, 2006). Social distance signals genetic proximity. In an intriguing study, DeBruine (2002) found that people trusted others more when the faces of these others were morphed to resemble their own (although there was no effect on trustworthiness). The effects of social distance are consistent with the theory of inclusive fitness (Hamilton, 1964). It would be odd if people felt as much benevolence or inequality aversion with respect to strangers as they feel with respect to their own children.

Social distance increases sharply when a categorical group boundary separates two individuals. Both trust and trustworthiness are stronger within groups than they are across group boundaries (Buchan, Croson, & Dawes, 2002; Glaeser, Laibson, Schenkman, & Souter, 2000; Tanis & Postmes, 2005). Lack of trust in an outgroup is to be expected if decisions to invest depend in part on social projection. It is a robust finding that people strongly project their own preferences and behaviors onto ingroup members, but hardly onto outgroup members (Robbins & Krueger, 2005). If individuals with cooperative preferences invest only inasmuch as they expect reciprocity (that is, inasmuch as they assume the other person also has prosocial preferences), they will regard trust in an outgroup member as riskier than trust in an ingroup member. Likewise, people may care less about their reputations in the eyes of an outgroup.

Even within a group, preferences for fairness are easily eroded by cues suggesting interpersonal competition. Garcia, Tor, and Gonzalez (2006) found that as pairs of players rank either very high or very low relative to all other players, they become more tolerant of inequality. They prefer a payoff that is larger than the other's payoff even when there is a larger alternative payoff for themselves that does not differ from the other's payoff. In other words, the same people can express competitive or cooperative preferences depending on where the exchange with an individual other is situated relative to the group. Even small changes in the context of the exchange can reveal "the naked expression of purely self-regarding behavior" (Smith, 1998, p. 16).

Conclusions

The evidence suggests that people tend to put their self-interest first, while being cognizant of relevant social norms. They apply these norms strategically, seeking to maximize their monetary rewards and to simultaneously enhance their self-image and reputation as a moral person. Ben-Ner and Putterman (2001, p. 529) put "a concern with how one is viewed by others" on a par with other biological preferences. This concern, they suggest, can be strong enough to create self-deception. "It is in the genetic interest of the individual to be perceived to be a cooperator, not necessarily to cooperate, (and so) we convince ourselves that we are indeed trustworthy, loyal, and moral, all the while on the look out to shade on reciprocity towards nonkin in favor of our immediate genetic interests" (Ben-Ner & Putterman, 2000, p. 94).

It might be hasty to consider social preferences irrelevant. Instead, it appears that stable individual differences are only part of the story, and difficult to isolate.[12] Benevolence and guilt-induced inequality aversion are naturally confounded with respect for social norms, and norm-adherence can be used to support reputations and positive self-images (Hoffman, McCabe, & Smith, 1996). For better or for worse, self-interest leaves a clearer footprint because no other social motive pulls behavior in the same direction. For the attainment of social harmony and efficient economic exchanges, a focus on social norms therefore seems to be especially promising (Bicchieri, 2006). Norms can be inculcated during socialization and activated by the presence of others. The perceived social distance of other individuals can be reduced by making common group memberships salient.

Given that norm adherence is flexible and often subordinated to self-interest, the question remains whether a utility model can be written that takes these additional variables into account. There are difficulties, however. If, for example, the social preference model were amended with weights that express social distance, the model's structure would still be the same. Trustees would still be choosing between reciprocating fully or not at all. Only the location of the difference point would vary. Arguably, an attempt could be made to formally capture concerns for positive self-images and desirable social reputations. These concerns would have be introduced as a measure of both individual and contextual differences.

For the sake of illustration, consider a trustee who seeks to balance self-interest against concerns of image and reputation in a particular context. Suppose the player's payoffs, $1 \leq X \leq 10$, are scaled logarithmically so that the monetary utilities are $U_{\mathrm{MONEY}} = \ln(X)$. This gain function captures the negatively accelerating psychophysics of money. Further suppose that the perceived moral benefits of other-regarding behavior

[12]The experimental study of the social preferences predates the individual-differences approach. Deutsch (1960) found that cooperation in the prisoner's dilemma was stronger when cooperative preferences, as compared with individualist or competitive preferences, were induced. In contrast, the individual differences that are the most useful predictors of strategic social exchange are precisely those that do not imply stable preferences regarding benevolence or fairness, such as Machiavellianism (Burks et al., 2003; Gunnthorsdottir, McCabe, & Smith, 2002).

are scaled inversely so that $U_{MORAL} = \ln(-X + 11)$. This loss function captures the idea that the difference between giving \$2 and giving \$1 is experienced as being larger than the difference between giving \$10 and \$9. The resulting subjective experience of $U_{MONEY} + U_{MORAL}$ is a quadratic function with a maximum at $X = 5.5$. Although it is a simple matter to model the trade-off between self-interest and moral concern this way, the experimental evidence reviewed earlier does not suggest that a general prescriptive model can be attained. Nor does it seem that a general descriptive model is likely to succeed because the individual and contextual differences are substantial.

Most of this discussion has been focused on the trustee because the implications of both, the social preference model and the norm adherence model, are easier to unpack. Yet, we note in closing that the motivations to enhance one's self-image or reputation are general psychological tendencies (Vignoles, Regalia, & Manzi, 2006). Inasmuch as a group or a society desires to increase trust, it might focus on reducing the trustors' perceptions of risk or on activating prosocial norms of giving. Cues suggesting a short social distance between trustor and trustee can stimulate investments. Trustors will project their own prosocial preferences (if they have any) onto trustees, they will feel compelled to adhere to norms, and they will expect that their norm-adherence be rewarded by enhanced reputations and self-images.

References

Adams, J. S. (1963). Toward an understanding of inequity. *Journal of Abnormal and Social Psychology*, *67*, 422–436.

Arrow, K. J. (1974). *The limits of organization*. New York: Norton.

Ben-Ner, A., & Putterman, L. (2000). On some implications of evolutionary psychology for the study of preferences and institutions. *Journal of Economic Behavior & Organization*, *43*, 91–99.

Ben-Ner, A., & Putterman, L. (2001). Trust and trustworthiness. *Boston University Law Review*, *81*, 523–551.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*, 122–142.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge, U.K.: Cambridge University Press.

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk, and betrayal. *Journal of Economic Behavior & Organization*, *55*, 467–484.

Bolton, G., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*, 166–193.

Buchan, N. R., Croson, R. T. A., & Dawes, R. M. (2002). Swift neighbors and persistent strangers: A cross-cultural investigation of trust and reciprocity in social exchange. *American Journal of Sociology*, *108*, 168–206.

Burks, S. V., Carpenter, J. P., & Verhoogen, E. (2003). Playing both roles in the trust game. *Journal of Economic Behavior & Organization*, *51*, 195–216.

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*, 7–42.

Cialdini, R. B. (2001). *Influence: Science and practice*. Boston, MA: Allyn and Bacon.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, *26*, 139–198.

Cox, J. C. (2003). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*, 260–281.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*, 260–281.

Cox, J. C., & Deck, C. A. (2005). On the nature of reciprocal motives. *Economic Inquiry*, *43*, 623–635.

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in a dictator game. *Organizational Behavior and Human Decision Processes*, *100*, 193–201.

Dana, J., Weber, R., & Kuang, J. X. (2007). Exploiting moral wriggle room: Behavior inconsistent with a preference for fair outcomes. *Economic Theory*, *33*, 67–80.

Dawes, R. M., McTavish, J., & Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology*, *35*, 1–11.

DeBruine, L. M. (2002). Facial resemblance and trust. *Proceedings of the Royal Society of London B*, *269*, 1307–1312.

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human Relations*, *13*, 123–139.

Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, *130*, 355–391.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behaviour. *Economic Inquiry*, *41*, 20–26.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*, 159–181.

Fiske, S. T., Cuddy, J. C., & Glick, P. (2006). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*, 77–83.

Garcia, S. M., Tor, A., & Gonzalez, R. (2006). Ranks and Rivals: A theory of competition. *Personality and Social Psychology Bulletin*, *32*, 970–982.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2005). *Moral sentiments and material interests*. Cambridge, MA: MIT Press.

Glaeser, E. L., Laibson, D. I., Schenkman, J. A., & Souter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics*, *115*, 811–846.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, *25*, 161–178.

Gunnthorsdottir, A., McCabe, K., & Smith, V. L. (2002). Using the Machiavellianism instrument to predict trustworthiness in a bargaining game. *Journal of Economic Psychology*, *23*, 49–66.

Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*, 245–256.

Hamilton, W. D. (1964). The genetical evolution of social behavior. *Journal of Theoretical Biology*, *7*, 1–52.

Hoffman, E., McCabe, K. A., & Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *American Economic Review*, *86*, 653–660.

Homans, G. C. (1958). Social behavior as exchange. *American Journal of Sociology*, *63*, 597–606.

Hume, D. (1739/1978). *A treatise of human nature* (L. A. Selby-Bigge, Ed.). Oxford: Clarendon Press.

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

Jones, B., & Rachlin, H. (2006). Social discounting. *Psychological Science*, *17*, 283–286.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology*, *89*, 899–913.

Kagel, J. H., Kim, C., & Moser, D. (1996). Fairness in ultimatum games with asymmetric information and asymmetric payoffs. *American Economic Review*, *76*, 728–741.

Kant, I. (1785/1964). *Groundwork of the metaphysic of morals* (H. J. Paton, Ed.). New York: Harper & Row.

Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations: A theory of interdependence*. New York: Wiley.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, *308*, 78–83.

Kramer, R. M. (2007). *Organizational trust*. New York: Oxford University Press.

Krueger, J. I. (2007). From social projection to social behavior. *European Review of Social Psychology*, *18*, 1–35.

Krueger, J. I., & Acevedo, M. (2005). Social projection and the psychology of choice. In M. D. Alicke, D. Dunning & J. I. Krueger (Eds.), *The self in social perception* (pp. 17–41). New York: Psychology Press.

Krueger, J. I., & Acevedo, M. (2007). Perceptions of self and other in the prisoner's dilemma: Outcome bias and evidential reasoning. *American Journal of Psychology*, *120*, 593–618.

Kuhlman, D. M., & Wimberley, D. C. (1976). Expectations of choice behavior held by cooperators, competitors, and individualists across four classes of experimental games. *Journal of Personality and Social Psychology*, *34*, 69–81.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Ethology and Human Biology*, *28*, 75–84.

Luhmann, L. (1988). Familiarity, confidence, trust: Problems and alternatives. In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (pp. 94–107). New York: Blackwell.

Malhotra, D. (2004). Trust and reciprocity decisions: The differing perspectives of trustors and trusted parties. *Organizational Behavior and Human Decision Processes*, *94*, 61–73.

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization*, *52*, 267–275.

Messick, D. M., & McClintock, C. G. (1968). Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology*, *1*, 1–25.

Mulder, L. B., van Dijk, E., De Cremer, D., & Wilke, H. A. M. (2006). Undermining trust and cooperation: The paradox of sanctioning in social dilemmas. *Journal of Experimental Social Psychology*, *42*, 147–162.

Murnighan, J. K., Oesch, J. M., & Pillutla, M. (2001). Player types and self-impression management in dictatorship games: Two experiments. *Games and Economic Behavior*, *37*, 388–414.

Peeters, G. (2002). From good and bad to can and must: Subjective necessity of acts associated with positively and negatively valued stimuli. *European Journal of Social Psychology*, *32*, 125–136.

Pillutla, M. M., Malhotra, D., & Murnighan, J. K. (2003). Attributions of trust and the calculus of reciprocity. *Journal of Experimental Social Psychology*, *39*, 448–455.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, *83*, 1281–1302.

Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review*, *9*, 32–47.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*, 466–469.

Smith, A. (1759/1976). *The theory of moral sentiments* (D. D. Raphael & A. L. Macfie, Eds.). Oxford: Clarendon Press.

Smith, A. (1776/1869). *An inquiry into the nature and causes of the wealth of nations* (J. E. T. Rogers, Ed.). Oxford: Clarendon Press.

Smith, V. L. (1998). The two faces of Adam Smith. *Southern Economic Journal*, *65*, 1–19.

Tanis, M., & Postmes, T. (2005). A social identity approach to trust: Interpersonal perception, group membership and trusting behavior. *European Journal of Social Psychology*, *35*, 413–424.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*, 183–206.

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*, 337–349.

Van Lange, P. A. M., & Liebrand, W. B. G. (1991). Social value orientation and intelligence: A test of the goal-prescribes-rationality principle. *European Journal of Social Psychology*, *21*, 273–292.

Vignoles, V. L., Regalia, C., & Manzi, C. (2006). Beyond self-esteem: Influence of multiple motives on identity construction. *Journal of Personality and Social Psychology*, *90*, 308–333.

Wojciszke, B. (2005). Morality and competence in person- and self-perception. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 16, pp. 155–188). New York: Taylor & Francis.

Theresa E. DiDonato's research interests include social cognition, interpersonal relationships, and the self. Her dissertation examines how the quality of close, personal relationships influences self-awareness and access to self-knowledge. Presently a Visiting Instructor at Wheaton College in Norton, MA, Theresa holds a BA from Wellesley College, and expects to graduate from Brown University with her doctorate degree in May 2008.

Joachim I. Krueger is Professor of Psychology and Human Development at Brown University. His research interests include self-perception, social stereotyping, and strategic reasoning in interpersonal and intergroup contexts. He has recently edited volumes on ''The self in social judgment'' (with M. Alicke & D. Dunning, Psychology Press, 2005) and ''Rationality and social responsibility'' (Psychology Press, in press) (http://research.brown.edu/research/profile.php?id=10378).

Adam L. Massey was born in Barstow, California. While in high school, he developed a strong interest in the sciences, particularly the mathematical sciences, and was admitted to Brown University in April of 2002. He studied mathematics at Brown, earning a Bachelor of Science in May 2006, before earning a Master of Arts in math from the University of California, Los Angeles, in June 2007. He is currently a PhD student at UCLA, where he focuses on algebra and number theory, and on the teaching of mathematics.