Evidential reasoning in the Prisoner's Dilemma

MELISSA ACEVEDO
University of Miami

JOACHIM I. KRUEGER
Brown University

Classic game theory considers defection to be the rational choice in the Prisoner's Dilemma. Although defection maximizes a player's gain regardless of the opponent's choice, many players cooperate. We suggest that cooperation can be explained in part by expectations of reciprocal behavior, that is, by the belief that players facing the same situation probably will come to the same decision. In Experiment 1, expectations of reciprocity were experimentally manipulated. As predicted, cooperation increased monotonically with these expectations. In Experiment 2, experimentally manipulated expectations of opponent cooperation were not associated with higher rates of cooperation. These findings are interpreted in light of a general model of payoff maximization, and implications for other models (e.g., social value orientation) are discussed.

In a social dilemma, a person faces a choice between two options. One is called cooperation, and the other is called defection. When one player in a two-person Prisoner's Dilemma (PD) game defects while the other cooperates, the former gains the most, and the latter gains the least. When both cooperate, both gain some but not as much as a unilateral defector does. When both defect, both gain little but more than a unilateral cooperator does. In other words, players collectively gain the most if they both cooperate, whereas they individually gain the most if they defect. The one-shot game casts this dilemma in its sharpest form. Here, players can neither shape each other's behavior through strategic rewards and punishments nor build reputations of being trustworthy. Classic game theory assumes that defection is the rational choice inasmuch as people are motivated by self-interest. A rational player is assumed to understand that defection dominates cooperation and therefore to prefer defection over cooperation no matter what. Defection is a "sure thing" (Savage, 1954); the opponent's behavior and payoff are irrelevant (Gibbons, 1992). If it does not matter what the opponent does, it is pointless to speculate what he or she might do.

Game theory does not ignore beliefs about the reasoning of others. The theory assumes that rational players expect their opponents to be rational. Being exposed to the rules of the game, players are assumed to realize that their opponents appraise these rules in the same rational (i.e., self-interested) way. They are assumed to know that their opponents know this, that their opponents know they know this, and so on. These shared meta-assumptions are the common psychological ground for the game (Nozick, 2001).[1] However, these assumptions are at odds with the claim that payoff dominance is a sufficient basis for rational choice. A player choosing the sure thing does not need to worry about the opponent's motivations or level of rationality.

Empirically, nearly half the choices in the PD are cooperative (Komorita & Parks, 1995; Sally, 1995). This is an anomalous result, given the assumption that defection is dominant (Dawes & Thaler, 1988), and the possibility that some players cooperate because of inattention or confusion (Andreoni, 1995) seems hardly sufficient to account for it. Other bases for rational choice must be considered.[2] Perhaps a player making assumptions about the opponent's rationality will not conclude that defection is the only rational option. Following Nozick (1993), we suggest that a player may engage in evidential reasoning. In this kind of reasoning, beliefs and expectations regarding the behavior of others matter a great deal, and thus the assumption of common ground becomes relevant. Assuming that their opponents are as rational as they themselves are, players can use their own choices to predict the choices of others. A player who is inclined to cooperate may think that the opponent is more likely to cooperate than to defect. A player who is inclined to defect is more likely to expect defection.

The mark of evidential reasoning is that the same player can alternate between choice options and thus between conflicting predictions before making a final decision (Krueger & Acevedo, 2005b). To illustrate, suppose a player believes that the opponent will reciprocate his or her own choice with a probability of .8. The expected value of cooperation, EV(c), is .8 times the payoff for mutual cooperation plus .2 times the payoff for unilateral cooperation. The expected value of defection, EV(d), is .8 times the payoff for mutual defection plus .2 times the payoff for unilateral defection. If EV(c) > EV(d), a self-interested player maximizes the expected value of the game by cooperating.

Figure 1 displays EV(c) − EV(d) for expectations of reciprocity from .5 to 1. The positive slopes mean that cooperation becomes more attractive as expectations of reciprocity increase. The differences in elevation between the three lines indicate that the level of expected reciprocity at which cooperation becomes more attractive than defection depends on the particular payoff matrix. In the sample matrices, the payoffs for

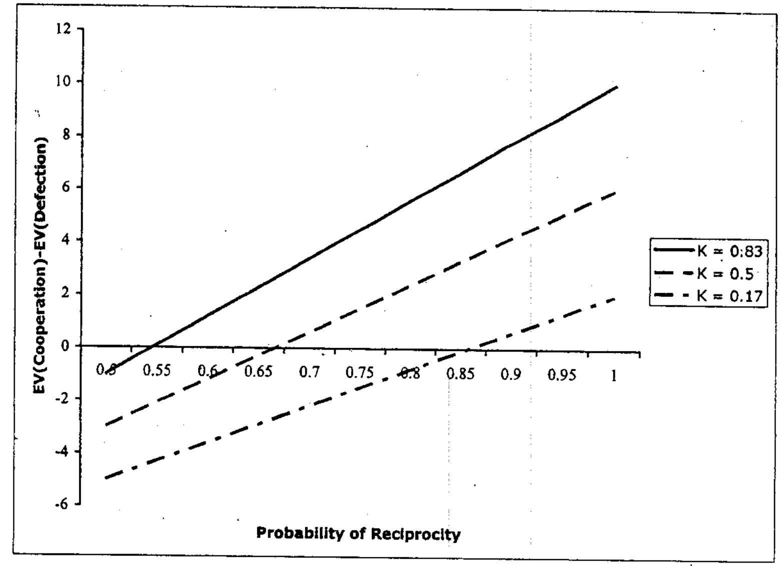

Figure 1. The expected reciprocity model

unilateral defection (T for *temptation*) and the payoffs for unilateral cooperation (S for *sucker*) are respectively 12 points and 0 points, whereas the payoffs of mutual cooperation (R for *reward*) and mutual defection (P for *punishment*) vary. The K ratio, which is $(R - P)/(T - S)$, determines how strong expectations of reciprocity need to be so that cooperation is more attractive than defection (Rapoport, 1967). When the probability of reciprocity is $1/(1 + K)$, the expected value of cooperation is equal to the expected value of defection (see proof in the *Appendix*).

We suggest that evidential reasoning has a psychological basis in processes of social projection. Social projection is the perception that others are more likely to be similar to the self than different. The analytical truth that most people are members of majorities ensures that projection increases the accuracy of social predictions (see Krueger, 1998, 2000b, for reviews).[3] This projection-induced accuracy benefit is larger to the extent that people know less about others. A player who is kept ignorant of the other player's motives and intentions is in precisely the kind of state that makes projection most effective. It is the ignorant player who has the most reason to believe that his or her own choice probably will turn out to be matched by the other.

An alternative psychological process that can create perceptions of similarity is self-assimilation. According to this view, people do not necessarily project their own behaviors to others. Under certain conditions, they may do the opposite. Starting from beliefs about which behaviors are common in the group, people may decide to fall in with the perceived majority (Karniol, 2003; Onorato & Turner, 2004). When applied to the PD, this idea suggests that a player may have a preconceived notion as to whether the opponent will cooperate or defect and may choose to match that choice. We now consider the implications of both hypotheses more closely.

Projection

In the PD, players' estimates of the other player's choice typically are collected after the player has indicated his or her own choice. The evidence shows that cooperators expect more cooperation than defectors do, and estimates are more accurate than they would be if players claimed to be agnostic about their opponents' behavior (Dawes, McTavish, & Shaklee, 1977; Kelley & Stahelski, 1970; Messé & Sivacek, 1979). Because these estimates are postdictions, they say little about the way in which projection can shape players' own decisions. In contrast, evidential reasoning suggests that antecedent projection (i.e., anchoring predictions of an opponent's choice on one's own intended choice) increases the willingness to cooperate.

On the surface, there is a difference between the postdiction and the prediction paradigm. In the postdiction paradigm, the cooperators and the defectors who make the estimates about others are different players. In the prediction paradigm, the same player is thought to make two conditional judgments before reaching a final decision. The player may ask, "What is the probability that my opponent will cooperate if I cooperate?" and "What is the probability that my opponent will cooperate if I defect?"

The distinction between postdiction and prediction has no statistical implications (Krueger & Acevedo, 2005b). If it is rational for two players with different preferences to make different predictions about the majority response, it is also rational for the same player to make different predictions when his or her own intended response changes. For example, when participants' own responses are manipulated by experimental design, their social predictions change accordingly (Clement & Krueger, 2002). The model of evidential reasoning assumes that similar changes occur (indeed, should occur) when a person merely considers alternative behavioral options. In the PD, a player needs to contemplate the implications of own cooperation and own defection before committing to one of these choices. Therefore, opponent cooperation should appear more probable when own cooperation is being considered than when own defection is being considered. Kay and Ross (2003) reported evidence supporting this idea. They primed participants to construe the PD as either a cooperative or a competitive game. This manipulation affected both the participants'

own behavioral intentions and the choices they expected other players to make.

Projection is a source of "cheap thought," allowing players to simulate what others will do based on what they themselves intend to do (Goldman, 2001; Orbell & Dawes, 1991). While they contemplate the available options, their own choices still hang in the balance. This dynamic shift from one potential option (and its corresponding prediction) to the other captures the intuition that players pass through a "Shall-I-or-shall-I-not" loop before crossing the decisional Rubicon. Expectations of reciprocity allow players to assess the expected values of cooperation and defection and to choose the more attractive option. Thus, our first hypothesis is that the probability of reciprocity predicts the rate of cooperation in a one-shot PD.

Across studies, higher K ratios elicit more cooperation (Dawes & Orbell, 1995; Komorita, 1976; Komorita, Sweeney, & Kravitz, 1980; Orbell & Dawes 1993). Parks, Sanna, and Berel (2001, p. 347) concluded that "the K value of a pay-off matrix indicates the rate of cooperation one would expect from a typical person." But what is it about a high K ratio that makes cooperation attractive? The dominance principle does not provide an answer because it mandates defection as long as the inequality of payoffs that defines the PD is satisfied. However, evidential reasoning does predict the effect of the K ratio. For a given expectation of reciprocity, a higher K ratio is more likely to yield $EV(c) > EV(d)$. Thus, we hypothesize that a high K ratio elicits more cooperation than a low K ratio (Experiment 1).

Assimilation

The alternative to the projection hypothesis is the idea that choices reflect players' efforts to reciprocate what they perceive to be the opponent's most probable choice. In its elementary form, the assimilation hypothesis states that cooperation increases with the perceived probability of opponent cooperation. Two lines of reasoning suggest that this might be so. First, the norm of reciprocity demands that people treat others the way they themselves have been treated (Gouldner, 1960). The norm of reciprocity makes cooperation the moral choice when opponent cooperation is evident or at least suspected. Second, assimilation is a successful strategy in repeated social dilemmas because it shapes the ongoing behaviors toward the optimum collective outcome (Axelrod & Hamilton, 1981; Rapoport, 1987). Although a strategy of reciprocity can be used only in a repeated PD, some players may generalize its effectiveness to one-shot dilemmas. As a result, they may assimilate their own behavior to what they think their opponent will do.

In the present experiments, we directly manipulated players' expectations of the probability with which the opponent would reciprocate their choice (Experiment 1) and the probability with which the opponent would cooperate (Experiment 2). The probability of reciprocity was a proxy for the psychological process of projection, and the probability of cooperation was a proxy for a stable assumption (i.e., stereotype) about what another person would choose to do.

EXPERIMENT 1

Participants played multiple rounds of the PD against a computerized opponent (see also Baker & Rachlin, 2001). During the course of the experiment, they received payoff matrices with varying K ratios. The probability of reciprocity was manipulated, and it was predicted that rates of cooperation would increase as a direct function of this probability. It was also predicted that a higher K ratio would elicit more cooperation. On control rounds, no probability information was introduced. On these rounds, cooperation was expected to be rare. The reason for this prediction was that participants playing against a computer cannot project their own intentions and therefore have no grounds for evidential reasoning. In other words, in the computerized PD we expected to see evidence of evidential reasoning when the probability of reciprocity was explicitly manipulated.

METHOD

Participants

Brown University undergraduates (29 men and 51 women) from lower-division psychology courses participated in partial fulfillment of a course requirement. The five participants who accumulated the highest number of points in the PD across experimental sessions each received a $50 bonus.

Procedure

Participants were recruited in groups of 5–10 for an experiment on decision making. Each participant was placed at a separate work station and was told that the experiment involved playing 18 rounds of an experimental game against a computer. Participants were informed that a $50 bonus would be awarded to the five participants who scored the most points in the game. The game was illustrated with a standard payoff matrix ($T = 12$, $R = 8$, $P = 4$, $S = 0$). To ensure a neutral instructional set, cooperation was called Option A, and defection was called Option B. Before the game started, participants were quizzed on its rules to ensure they understood the implications of their choices. The experimenter cleared up any misunderstandings or questions participants had before the game.

On nine rounds, participants learned that they would receive the probability with which their own choice would be reciprocated. A probability of .5 indicated the lack of reciprocity, a probability of .75 indicated an intermediate level of reciprocity, and a probability of 1.0 indicated complete reciprocity. On six rounds, participants were informed that they would know their opponent's choice before making their own. On the remaining three rounds, they were not to receive any information. Finally, participants were told that the payoff matrices would vary from round to round. The values for T and S would always be 12 and 0, respectively, and the paired values for R and P would be 11 and 1 ($K = .83$), 9 and 3 ($K = .50$), or 7 and 5 ($K = .17$).

The experiment had a 6 (decision rule: $p = .5$, .75, or 1, no information, certain cooperation, and certain defection) $\times$ 3 ($K = .83$, .50, .17) repeated-measures design. The dependent measure was the choice between Option A (cooperation) and Option B (defection). Instructions were presented and data were collected on Macintosh computers, using Superlab (Haxby, Parasuraman, Lalonde, & Abbound, 1993). Responses were coded as 1 for defection and as 2 for cooperation and summed for each payoff matrix to provide a measure of cooperation. The order of the 18 rounds was randomly determined for each participant. At the end of the session, participants were debriefed and probed for suspicion. Bonus money was mailed to winning participants at the conclusion of the experiment.

RESULTS

Probability of reciprocity

The two main hypotheses that motivated this study were that rates of cooperation would increase with the probability of reciprocity and with the K ratio of the payoff matrix. The findings supported both hypotheses (Figure 2a). Rates of cooperation were low when there were no expectations of reciprocity (23%), moderate for intermediate expectations (66%), and high for high expectations (93%). A pair of Wilcoxon's tests for matched samples confirmed that the moderate rate of cooperation was significantly higher than the low rate, $z = 6.63$, $p < .001$, and significantly lower than the high rate, $z = 5.17$, $p < .001$. This rise of cooperation across levels of expected reciprocity was observed for each individual matrix, McNemar's $Q(2) = 63.51$ for $K = .17$, 81.34 for $K = .5$, and 77.23 for $K = .83$, all ps $< .01$.

Also as expected, rates of cooperation increased with the K ratio of the payoffs, but this effect was limited to the intermediate level of expected reciprocity, $Q(2) = 12.81$, $p < .01$. When expected reciprocity was at its maximum, the K ratio was irrelevant, $Q(2) = 3.25$, *ns;* when there were no expectations of reciprocity, a small but significant effect of payoff emerged, $Q(2) = 4.21$, $p < .02$, but it did not represent any meaningful order (i.e., that of the K ratio).

Together, the effects of expected reciprocity and the K ratio constituted the football-shaped pattern shown in Figure 2a. By contrast, the pattern of theoretical predictions displayed in Figure 1 involved linear effects of expected reciprocity and the K ratio. This difference in pattern can be attributed to the fact that Figure 1 displays the interval-scaled differences between two expected values (i.e., $EV[c] - EV[d]$), whereas Figure 2a

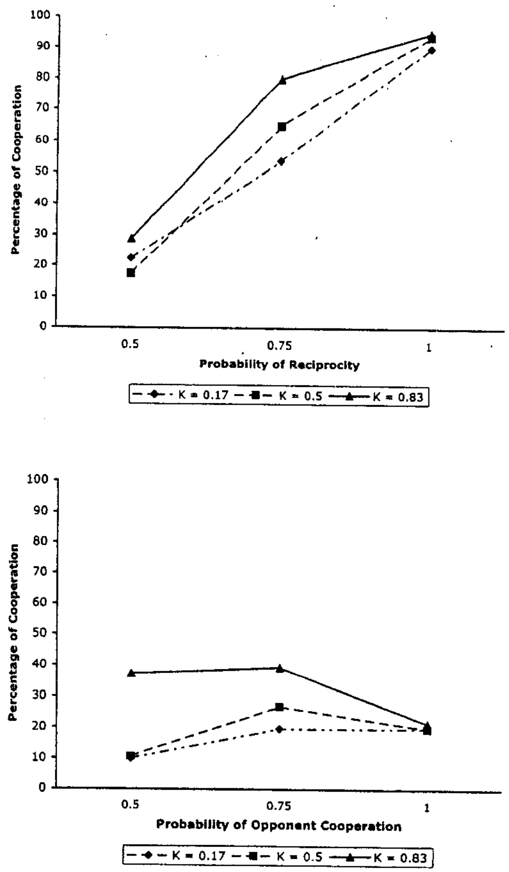

Figure 2. The effects of (a) probability of reciprocity and (b) probability of opponent cooperation on participant cooperation

displays categorical choices as proportions. A strict extrapolation from the theoretical model would predict an even more strongly inflated football. For intermediate expectations of reciprocity, the K ratio ought to make an all-or-none difference; the rate of cooperation should approach 100% for $K = .83$ and $K = .50$, and it should approach 0% for $K = .17$. To account for the contraction of the empirical data relative to the strict theoretical model, we later sketch a descriptive judgment model that takes perceivers' biases and computational errors into account.

Control conditions

There were three control conditions. When participants were led to expect opponent cooperation or defection with certainty, or when they received no information, their own choices were not diagnostic of opponent choices. As expected, few participants cooperated when anticipating cooperation (7%) or when anticipating defection (5%). When they had no information, they were more likely to cooperate (21%; $z = 2.06$, $p = .04$ for the comparison with the certain cooperation condition, and $z = 4.44$, $p < .001$ for the comparison with the certain defection condition). These differences emerged for each K ratio, $Q(2) = 5.09$, 6.63, and 20.18, all $ps < .05$, respectively, for the ratios of .17, .5, and .83; see Table 1).

Unexpectedly, cooperation increased with the K ratio in the no-information condition, $Q(2) = 14.63$, $p < .001$, and in the certain cooperation condition, $Q(2) = 8.44$, $p < .01$. The K ratio was irrelevant only in the certain defection condition, $Q(2) = 0$. These effects could be explained neither by game theory nor by the model of expected reciprocity.

DISCUSSION

In support of our main hypothesis, cooperation rates rose sharply with increases in expectations of reciprocity. When expectations were of intermediate strength, a high K ratio elicited the most cooperation. A simple multiplicative model, combining the probability of reciprocity with the K ratio, predicted cooperation rates across conditions. Across the nine experimental conditions, the differences between the expected values of cooperation and the expected values of defection were highly correlated with the rates of cooperation, $r = .84$, $p = .01$. By itself, the K ratio was a poor predictor of cooperation, $r = .17$, $p = .48$. Expected reciprocity thus appeared to be necessary for substantial cooperation to emerge. Cooperation rates were low when no information about opponent behavior was given or when players were assured that they would encounter cooperation or defection.

The rate of 20% cooperation emerged as a baseline of choice under uncertainty. This rate was observed when the probability of reciprocity was .5 and when no information was available about computer choice. In this experiment, the no-information condition did not reflect the situation a player faces with a human opponent. Social projection was disabled

Table 1. Effects of payoff and decision rule on participant cooperation, Experiment 1

	K ratio		
Computer's decision rule	.17	.5	.83
No information	12.5%	16.3%	31.3%
Chooses A (cooperation)	7.5%	12.5%	20.0%
Chooses B (defection)	5.0 %	5.0%	5.0%

because whichever choice the player may have been contemplating, his or her processes of deliberation were not projectable to a computer. Any simulation heuristic or theory of mind applicable to humans was of little value. Relative to this baseline, cooperation was reduced when opponent choice was known. However, even when cooperation was certain, some players cooperated, although they should have recognized the availability of a free ride. Their failure to do so cannot be reconciled with game theory or with any theory assuming a generalized bent toward altruism.

The hypothesis of evidential reasoning suggests that choices in the PD are sensitive to expectations of reciprocity. Participants need to generate their own expectations, and social projection is a mechanism by which they can do this. Evidence for the alternative process of assimilation was weak. Participants did not appear to align their choices with what they were led to expect from the opponent. In the present design, it was not possible to compare projection and assimilation effects directly. Projection was simulated by variations in the expectations of reciprocity, whereas assimilation was simulated by telling players what the opponent would do. In Experiment 2 we varied the probability of opponent cooperation directly to facilitate comparisons across studies.

EXPERIMENT 2

In this experiment, players received the probability with which the opponent would cooperate. This information made explicit the game theoretic assumption that the probability of cooperation is independent of the player's own contemplated choice. Thus, this method ensured that players could not expect any systematic reciprocation of their choices. Both game theory and the model of evidential reasoning predicted that rates of cooperation would be uniformly low. In contrast, the assimilation hypothesis predicted that players would use the provided probability of cooperation as a cue toward their own decision to cooperate. In Experiment 1, the K ratio of payoffs was modestly associated with the rate of cooperation when cooperation is certain. We wanted to see whether this effect would replicate for varying probabilities of opponent cooperation.

METHOD

Participants

Brown University undergraduates (18 men and 38 women) from introductory psychology courses participated for extra course credit. The five participants who accumulated the most points each received a $50 bonus.

Procedure and design

The procedure was similar to that of Experiment 1. Participants played 12 rounds of a one-shot PD against the computer. On nine rounds, they were given the probability of opponent cooperation, namely .5, .75, or 1. On the three remaining rounds, there was no information. The order of rounds was randomized for each participant. The computer recorded the choices between option A (cooperation) and option B (defection).

RESULTS

The assimilation hypothesis stated that cooperation would increase with the probability of opponent cooperation. The findings, which are displayed in Figure 2b, showed that this was not so. Regardless of whether expectations of cooperation were low (p = .5, 21%), intermediate (p = .75, 29%), or high (p = 1, 20%), few players cooperated. The difference between the low and the intermediate probability of opponent cooperation was not significant, $z = 1.04$, $p = .19$; the difference between the intermediate and the high probability, though significant, $z = 2.27$, $p = .01$, indicated a decrease rather than an increase in cooperation. Across the three probability levels, cooperation did not vary significantly for $K = .17$, $Q(2) = .44$. For $K = .5$ and .83, there was significant variation in the rate of cooperation, $Q(2) = 6.42$ and 6.07, both ps < .05, but there were no linear increases.

When the effect of the K ratio was examined, significant differences emerged when p(cooperation) = .5, $Q(2) = 21.00$, $p < .01$, and when p(cooperation) = .75, $Q(2) = 10.33$, $p < .01$. Only for the latter effect did levels of cooperation correspond to differences in the K ratio. When cooperation was certain, there was no effect, $Q(2) = 2.14$, *ns*. The results from the no-information condition replicated the results from Experiment 1. Cooperation was lower for $K = .17$ (4.5%) than for $K = .5$ (14.3%) or $K = .83$ (13.4%), $Q(2) = 12.33$, $p < .01$.

DISCUSSION

As expected, rates of cooperation did not increase with the probability of opponent cooperation. Instead, they remained close to the baseline of 20% even when the probability of opponent cooperation was high. Some effects of differences in the payoff structure were observed, however. These effects were small, but the K ratio predicted the rate of cooperation across the nine experimental conditions in which the probability of opponent cooperation was provided, $r = .67$, $p = .06$. This is consistent with our quantitative model, which states that the correlation between

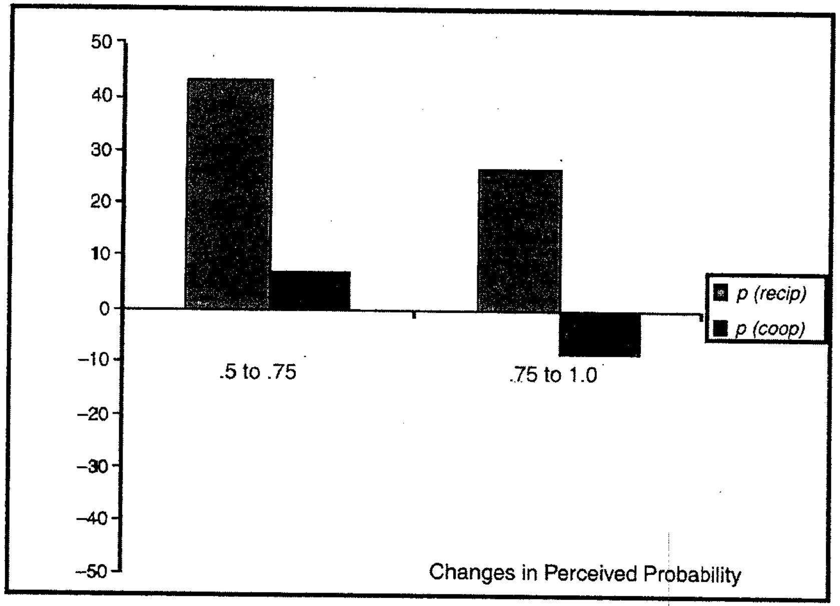

Figure 3. Differences in percentage of cooperation as a function of information about the probability of reciprocity, p(recip), and the probability of cooperation, p(coop)

EV(c) − EV(d) and cooperation is identical to the correlation between the K ratio and cooperation when the opponent's choice is statistically independent of the player's own choice.

Although game theory is consistent with the low rates of cooperation in Experiment 2, it cannot account for the increases observed in Experiment 1. By contrast, the model of evidential reasoning is consistent with the findings from both studies. To allow direct comparisons, we plotted the differences in cooperation between paired probability conditions. Figure 3 shows that when players regarded their own choices as diagnostic of opponent choices, cooperation drastically increased with higher levels of expected reciprocity (Experiment 1). When the same probability values referred to the probability of opponent cooperation, changes in participant cooperation were minimal (Experiment 2).

GENERAL DISCUSSION

The PD poses the puzzle that "by defecting and doing what seems individually reasonable and rational, people end up doing less well than they would have if they had acted unreasonably or irrationally" (Dawes & Messick, 2000, p. 111). A resolution of this puzzle entails a reconciliation

of prescriptive norms with observed behavior. By highlighting the role of evidential reasoning, we hope to contribute to a psychological game theory that may ultimately accomplish this (Colman, 2003). Commenting on Colman's proposal, Monterosso and Ainslee (2003, p. 171) noted that of "particular importance in explaining otherwise anomalous play is the subject's use of her own inclinations and behavior as test cases that inform her expectations regarding what others will do." In a related comment, Anatol Rapoport (2003, p. 173) argued that the "frequent failures of rational choice theory to prescribe acceptable actions in gambling or game-like situations can be traced to two often tacitly implied assumptions, namely, rejection of so called evidential decision theory . . . and independence of decisions of individual players." The model of evidential reasoning proposed here breaks with these tacit assumptions without abandoning the view that most people are motivated by rational self-interest. How well, then, does the model account for other empirical data?

Empirical fit

Dawes (1989) used Laplace's (1814) principle of insufficient reason to model social projection. Under the assumption of uniform priors, Bayes's rule states that one's own behavior will be reciprocated with a probability of two thirds. Although Bayesian predictions can be obtained from any set of prior probabilities, uniform priors are psychologically attractive because they capture a player's agnostic state of mind when confronted with the game. For an intermediate set of payoffs (i.e., $T = 12$, $R = 8$, $P = 4$, $S = 0$), cooperation is the most attractive choice when projection rises above .7 (Figure 1). The average empirical rate of cooperation of about 40% (Sally, 1995) is consistent with an average projection rate somewhat below .7.

In the PD, individual differences in projection produce variability in players' choices. According to a strict expected-value model, however, all participants should either cooperate or defect when the necessary information about probabilities and values is provided. When the probability of reciprocity is experimentally controlled, individual differences in projection are eliminated. Yet, as the findings displayed in Figure 2a show, individual differences in choice emerged as a function of payoff, especially when the probability of reciprocity was intermediate.

To account for this variability, we propose a psychologically plausible extension of the expected-value model. Recall that there was more cooperation in Experiment 1 than the strict model anticipated. According to that model, participants should have uniformly cooperated in five conditions (when $p = 1$, or when $p = .75$ for $K = .50$ or .83) and uniformly defected in four conditions (when $p = .5$, or when $p = .75$ for $K = .17$). This distribution of choices would have amounted to a total percentage of cooperation of 56%. The corresponding empirical rate was 61%. We think it plausible that there was a bias toward cooperation in the magnitude of

5%. This is consistent with Shafir and Tversky's (1992) finding that a few players cooperate no matter what (see also the current Experiment 2). Thus, choices in the PD can be modeled as $[EV(c) - EV(d)] + \text{Bias} + \text{Error}$, where players opt to cooperate if their calculations return a positive value.

The bias term may be positive for a number of reasons. For example, if participants labeled Option A "cooperation," the label itself could increase the appeal of this option because of its socially desirable connotations. As Fantino and Stolarz-Fantino (2003, p. 160) noted, people "have been taught since childhood to be unselfish and cooperative, [so why should these lessons] not 'intrude' on the decisions subjects make in the laboratory" (see also Rachlin, 2002). Alternatively, participants may simply favor whichever option they look at first without realizing why (Nisbett & Wilson, 1977; see also Schelling, 1960, for the psychology of focal points in coordination games).

The error term recognizes people's fallibility in calculations of expected values. To the extent that any random error is added to the computed difference between $EV(c)$ and $EV(d)$, some empirical choices will differ from the theoretically predicted ones, and the resulting rates of cooperation will contract toward the middle of the scale. This effect is large inasmuch as computations are difficult. In the present design, it was the intermediate probability of reciprocity that made the greatest demands on participants' ability and willingness to calculate.

Our experimental paradigm had two unique features: Participants played against an inanimate opponent, and the probabilities of reciprocity were controlled. The former feature minimized the relevance of emotions, such as hope, sympathy, or guilt. The latter treated expectations of reciprocity (or cooperation) as an independent rather than a dependent variable. Arguably, this method transformed the player's task from a game theoretic one to a decision theoretic one (Baker & Rachlin, 2002). In a game theoretic problem, players face ambiguity in the sense that they do not even know the probabilities with which other players will choose from the available response options. In a decision theoretic problem, relevant probabilities are available, and the task of research is to determine how people use them.

The question remains of whether this transformation is useful when trying to understand why people cooperate. We believe that it is because it shows that the PD is underspecified in its canonical game theoretic representation. The dominance principle holds that defection is the rational choice no matter what. Therefore, any probability estimates attached to opponent's expected behavior are irrelevant. The model of evidential reasoning makes room for a lawful process generating probability estimates. When the generation of such probabilities is derived from a psychological

process (projection) and reconciled with a normative theory of choice (maximization of expected value), it is reasonable to control the relevant probability information experimentally.

By its very design, this experimental method cannot shed light on the question of how much evidential reasoning occurs when anonymous human players are matched up in a game. The answer to this question requires additional evidence from correlational studies. We know of four empirical studies in which individual differences in social projection predicted cooperative choices. Participants in three of these studies estimated the probability that their favored political party would win a national election if they themselves voted and the probability of winning if they did not vote (Acevedo & Krueger, 2004; Krueger & DiDonato, 2005; Quattrone & Tversky, 1984). The difference between these two estimates served as an index of projection. Some participants, more than others, expected that their own choices between voting and abstention were diagnostic of the choices made by other supporters of the same party. Consistent with evidential reasoning, the index of projection was positively correlated in each study with the reported intention to vote. The weighted meta-analytic effect size was $r = .24$ (see Lipsey & Wilson, 2001).

In a fourth study, participants described themselves by rating their own personalities on a series of trait adjectives and then estimated the percentage of students who would endorse each trait as being self-descriptive (Krueger & Acevedo, 2005a). The correlation between self-ratings and consensus estimates, computed across traits for each participant, served as an idiographic index of social projection. This index predicted the likelihood of cooperation in a PD, $r = .19$.

Rationality revised

One criticism of evidential reasoning is that when two people act independently, they cannot influence each other's choices, however desirable that may be. A player in a one-shot PD cannot induce another player to cooperate merely by his or her own cooperative choice. Yet a player is justified to assume that his or her choice is more likely to be reciprocated than not.

Participants in Experiment 1 had causal power over the computer's choices. The probability of reciprocity was cast as a conditional probability of the computer's choice given the participant's choice. Note that when players do not possess explicit information regarding this probability, they can reason evidentially by first considering the inverse conditional probability, which is the probability that their own ultimate choice probably will be whichever choice is most common in the population. Although inverse conditional probabilities need not be identical, the ratio of correct predictions over wrong predictions is the same regardless of whether predictions are conditioned on the player's own choices or the opponent's choices. Predictions are correct if a player cooperates and the opponent also cooperates as expected or if a player defects and the opponent defects as expected. Predictions are wrong if a player cooperates and the opponent unexpectedly defects or if a player defects and the opponent unexpectedly cooperates. Evidential reasoning works because $[p(R)p(P)]/[p(S)p(T)] > 1$.

The crux of evidential reasoning is that it applies equally to behavior that has occurred and to behavior that is still under consideration. Before committing to a choice, a person can (and should) generate different predictions depending on which behavioral alternative is being considered. It is then tempting to think that these divergent probability estimates should cancel each other out. After all, if two estimates differ, at least one must be wrong. However, the same logic suggests that divergent estimates reached by two different players should also cancel each other out. If they differ, at least one must be wrong. The rationality of a single player to generate divergent probability estimates can be denied only if the rationality of different players to generate different estimates is also denied. To deny the former but not the latter is not a coherent position. Strict adherence to the dominance principle denies players the accuracy benefit resulting from projection when choices have become real (see Krueger & Acevedo, 2005b, for a full discussion).

Thus far, we have considered self-interest as the only relevant motivation. Neither the dominance principle, as stated by game theory, nor social projection, as assumed by the model of evidential reasoning, requires players to wonder what their opponents might want. However, both theories assume that players are self-interested and that they attribute self-interest to each other. In this sense, both theories admit the notion of projection. From the perspective of game theory, the attribution of self-interest to the opponent makes defection all the more compelling because the opponent too must be expected to defect. From the evidential perspective, however, cooperation may be more compelling. Examining the payoff matrix, a player will notice that his or her hypothetical changes of choice have a greater effect on the opponent's outcome than on his or her own outcome. In the standard matrix (see intermediate line in Figure 1), a switch from defection to cooperation increases the opponent's payoff by eight points, whereas it decreases own payoff by four points. With the power to reward and punish being greater than the power to pursue self-interest, players can maximize their own self-interest indirectly by acting to maximize the interest of the other and by assuming that the other will do likewise.

If players discover this evidential path, their cooperative choices are both rational and ethical. As Nozick (2001, p. 243) observed, "Ethical

beliefs coordinate our actions, each to each other's, so as to achieve mutually beneficial action." By attenuating the conflict between rational thought and ethical action, social projection defuses the intrapsychic conflict between egotism and altruism. Instead of yielding to one motive or another, people can engage in mutually rewarding social exchange (Kelley & Thibaut, 1978; Kiyonari, Tanida, & Yamagishi, 2000). When these exchanges occur between members of the same group, they are particularly effective (Dion, 1973; Yamagishi & Kiyonari, 2000). We assume that this is so because projection, and thereby expectations of reciprocity, is directed primarily to social ingroups rather than to outgroups (Clement & Krueger, 2002; Robbins & Krueger, 2005).

Prisoners transformed

Although game theory and the model of evidential reasoning emphasize self-interest as the motive of choice, other motives may come into play. For example, theories of social value orientation assume that some players care not only about their own payoffs but also about the payoffs of others (Messick & McClintock, 1968). We use Van Lange's (1999, 2000) integrative model to explore the implications of evidential reasoning for players who are less egocentric than we have assumed thus far.

Theories of social value orientation suggest that there are three types of player. The individualist is interested only in maximizing own payoffs, and the competitor is interested in maximizing the difference between his or her own payoff and the other player's payoff. Both are drawn toward defection regardless of what the opponent does. We will not consider them further. The third type is the prosocial player, who seeks to maximize joint payoffs while being leery of differences between the two payoffs. This concern about fairness is widely held among humans (Krueger, 2000a) and nonhuman primates (Brosnan & de Waal, 2003).

With this set of preferences, prosocial players face an effective payoff matrix that is different from the given matrix. They transform the given payoffs by summing their own and the weighted other player's payoffs and by subtracting the weighted difference between the two. Suppose the weight accorded to own payoff is 1 and the weight w accorded to the other player's payoff and to the difference term ranges between 0 and 1. Then, the effective payoffs are

$$
\begin{aligned}
T' &= T + wS - w(T - S)\\
R' &= R + wR - w(R - R)\\
P' &= P + wP - w(P - P)\\
S' &= S + wT - w(T - S).
\end{aligned}
$$

The top panel of Figure 4 shows an objective matrix with $K = .5$. The center panel shows a subjective matrix for $w = .5$.[4] The transformation of

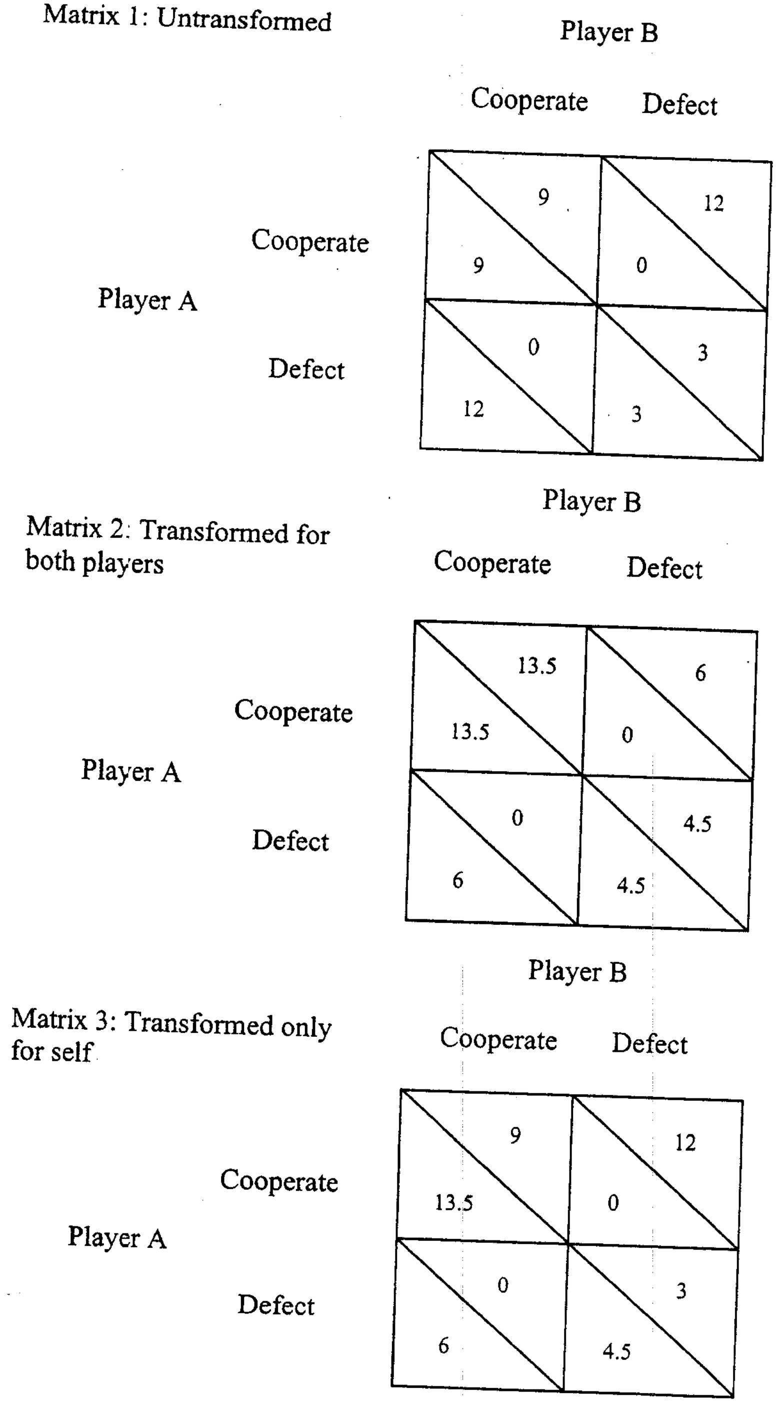

Figure 4. Prisoner's Dilemma payoff matrix and two transformations based on prosocial preferences

both players' payoffs indicates that the prosocial player assumes that the other player also holds prosocial preferences (Kuhlman & Wimberley, 1976). However, this projective inference is tempered by any self-enhancing expectation that others are less fair-minded than the self (Messick, Bloom, Boldizar, & Samuelson, 1985). Even prosocials may not believe that their opponents transform the matrix as much as they themselves do.

Once the payoffs are transformed, prosocial players find themselves in a Stag Hunt. This game, which is also known as an Assurance Game, has two equilibria. If Player B cooperates, Player A is better off cooperating; if Player B defects, Player A is better off defecting. Although players may prefer to coordinate their choice on mutual cooperation, they must fear that the other loses confidence in successful coordination and defects to secure a sure but small gain. There no longer is a dominating choice. Following the mandate of game theory, players need to resort to the minimax principle (von Neumann & Morgenstern, 1944), which advises them to choose whichever option results in the higher payoff assuming that the opponent does not choose as hoped. In a Stag Hunt (as in the PD), this is defection.

In contrast, evidential reasoning suggests that a Stag Hunt will elicit more cooperation than a PD because even for low levels of projection the expected value of cooperation is higher than the expected value of defection. Recall that a player is indifferent if the probability of reciprocity is $1/(1 + K)$. The K ratio for the transformed matrix is 1.5, which means that cooperation is attractive if p(reciprocity) > .4. The weight w accorded to the other player's payoff (and to the difference between own and other's payoffs) is critical. If $w \leq (R - T)/ (2S - T - R)$, the transformed matrix remains a PD. If $w = 1$, however, the game turns into a symmetrical coordination game, and projection is no longer useful. Regardless of the perceived probability of reciprocity, the expected value of cooperation is higher than the expected value of defection.

Now consider the bottom matrix. Here the prosocial Player A does not project his or her subjective payoffs to Player B. Seeing himself or herself as a participant in a Stag Hunt, this player assumes that the other player considers the game a PD. This is an odd situation because it abandons the standard symmetry and hence the assumption of common ground. The prosocial player now has two options. If the other player is assumed to endorse the dominance principle and defect, this player should also be tempted to defect. This strategy reflects the observation that even prosocial players are quick to assimilate to others' defection. A return to defection need not result from "a strong desire to restore equality" (De Cremer & Van Lange, 2001, p. 15) but simply from the prudent goal not to be suckered.

Alternatively, prosocial players may look to evidential reasoning to disambiguate the game. If they approach this mixed game only from their own construal of it as a Stag Hunt, they might cooperate too enthusiastically. If they acknowledge that their opponents continue to regard the game as a PD, it should be clear that those other players require a higher level of projection to find cooperation attractive. This realization yields a higher risk of being suckered than does a game that is believed to be treated as a Stag Hunt by both players. To reduce this risk, players can ask whether their own evidential reasoning would lead them to cooperate if they left their own payoffs untransformed. Ironically, this maneuver makes the player's own payoff transformation unnecessary.

Although theories of social value orientation have uncovered important individual differences in people's preference structure, they have not yet isolated the decision mechanisms that translate a preference for cooperation into actual choice. Our analysis suggests that although payoff transformations may eliminate the dominance principle as a viable decision rule, they do not eliminate the need for evidential reasoning; they reinforce it.

Why did the chicken cross the road?

We have suggested that evidential reasoning can be rationally applied to two different games (i.e., the PD and the Stag Hunt), whereas the dominance principle works only for games with a single Nash equilibrium (here, the PD). The greater generality of evidential reasoning makes it attractive as a rational approach to game situations. This argument is appealing to the extent that it holds for a broad range of games. To demonstrate the greater generalizability of evidential reasoning, we now consider the Chicken Game.

Along with the PD and the Stag Hunt, the Chicken Game, which is also known as the Hawk–Dove Game or the Snowdrift Game, ranks among the three most popular and intriguing experimental games (Kollock, 1998). It is characterized by a payoff matrix that satisfies the inequality of $T > R > S > P$. The game is superficially similar to the PD, yet the critical difference is that being suckered is not as bad as mutual punishment. The reversal of the ranks of the S and P payoffs means that there are two Nash equilibria and thus no dominating choice. Players do best if they manage to avoid making the same choice. In this sense, the Chicken Game is the inverse of the Stag Hunt. Nonetheless, collective cooperators do better than collective defectors.[5]

In the absence of a dominating choice, game theory must again resort to the minimax rule, which, in contrast to the PD and the Stag Hunt, returns the advice that it is best to cooperate. As in the other two games, evidential reasoning does not make a categorical recommendation. Instead, it suggests that one should use expectations of reciprocity to compute expected

values for both cooperation and defection and should choose whichever option yields the higher values. Figure 5 (*top*) shows an objective payoff matrix. For ease of comparison with the PD matrix in Figure 4 (*top*), the K value is set at .5. Because the point of indecision lies at $p(\text{reciprocity}) = 1/(1 + K)$ for both games, players in the PD and in the Chicken Game should cooperate if their expectations of reciprocity exceed .67.

The bottom panel of Figure 5 displays a payoff matrix that has been transformed in the same way as the matrix in the center of Figure 4. Whereas the transformation changes the PD into a Stag Hunt, it does not change the Chicken Game. It only changes the K value from .5 to 1.5, rendering even a low level of projective reasoning, $p(\text{reciprocity} > .4)$, sufficient to induce cooperation. Consistent with this analysis, rates

Matrix 1: Untransformed

Player A \ Player B	Cooperate	Defect
Cooperate	4.5, 4.5	3, 12
Defect	12, 3	0, 0

Matrix 2: Transformed for both players

Player A \ Player B	Cooperate	Defect
Cooperate	6.75, 6.75	4.5, 9
Defect	9, 4.5	0, 0

Figure 5. Untransformed and transformed payoff matrices for a game of Chicken

of cooperation tend to be higher in the Chicken Game than in the PD (Hauert & Doebell, 2004).

Optional play

As a final reflection on the general utility of the evidential reasoning model, we should note the similarities and differences between our model and the cognitive miser theory of cooperation (Dawes & Orbell, 1995; Orbell & Dawes, 1991, 1993). Like our model, the cognitive miser theory considers social projection critical for high rates of cooperation to be attained. The two key features in tests of the cognitive miser approach are that any player can veto the game after inspecting the payoff matrix and that the P and S payoffs constitute actual losses of money. The two key psychological assumptions are that players formulate an intention to cooperate or to defect upon examining the matrix and that only then do they project their own intention onto the other player. It follows—and it was empirically demonstrated—that players intending to defect are most likely to veto the game, evidently because they anticipate a loss due to mutual defection.

Our model does not require players to make a choice before projecting it to the other. Instead, it assumes that cooperation results from a comparison of two expected values, EV(c) − EV(d), where both depend in part on the choice that is projected to the other. Our model is more parsimonious in that it does not need to solve the problem of how a player can formulate an initial behavioral intention before considering what the other might do. The cognitive miser theory is mute on this point. It assumes only that initial intentions do not reflect the kinds of personality dispositions the theory of social value orientation refers to.

Conclusions

Game theory makes different categorical demands for different games, and social values can alter the structure of some games. In contrast, the model of evidential reasoning is coherent and uniform in its predictions. Rates of cooperation depend on the average level of projection present in a group of players and on the K value of the payoff matrix. Stronger projection and higher K values should lead to more cooperation.

The PD may present less of a dilemma to people who play it than to researchers who study it. Projecting players not only make more accurate social predictions than agnostic players do, they also end up doing well by cooperating. Because it provides no feedback about the choices of others, the one-shot PD is commonly understood as a qualitatively different situation than its iterative counterpart. In iterated games, players can learn about the strategies of others and try to shape their behavior (Komorita, Parks, & Hulbert, 1992). Reciprocal behavior, as exemplified by the tit-for-tat strategy, increases cooperation and thereby the individual and the

collective good (Axelrod & Hamilton, 1981). Tit-for-tat induces players to expect reciprocity. The more they honor this expectation in their own decisions, the better they do.

The tit-for-tat strategy makes two additional assumptions. One is that the first move must be cooperative. The present analysis suggests that this initial act of cooperation need not be made on faith. Inasmuch as the first round of an iterated game is equivalent to one-shot PD, evidential reasoning may be sufficient to warrant cooperation. As the game progresses, empirical expectations of reciprocity can then form and be updated online. As in the one-shot game, players who reciprocate and expect reciprocity will do best. The other assumption is that players who are committed to a tit-for-tat strategy are immune to the temptations of backward induction. Players who reason by backward induction notice that on the last round of the game, defection is the dominating strategy. If one must defect on the last round and expect the opponent to do likewise, the penultimate round effectively becomes a last round as well, and the same inference applies. Moving this reasoning backward, cooperative reciprocity never takes hold (Luce & Raiffa, 1957). Again, evidential reasoning suggests a way out because players need no longer assume that everyone will necessarily honor the mandate of dominance. On the last round, as on any others, players can reason that their own choices predict what others will do.

Appendix. Finding the level of projection at which the expected value of cooperation is equal to the expected value of defection, where p stands for $p(M)$

$$\mathrm{EV}(c) - \mathrm{EV}(d) = pR + (1-p)S - (1-p)T - pP$$
$$\mathrm{EV}(c) - \mathrm{EV}(d) = 0 \text{ if } pR + (1-p)S = (1-p)T + pP, \text{ that is, if}$$
$$pR - pP = (1-p)T - (1-p)S, \text{ that is, if}$$
$$p(R-P) = (1-p)(T-S), \text{ that is, if}$$
$$(R-P)/(T-S) = K = (1-p)/p.$$

Solving for p,

$$1 - p = Kp,$$
$$p = 1 - Kp,$$
$$1 = 1/p - K,$$
$$1 + K = 1/p,$$
$$1/(1+K) = p.$$

Notes

We thank Ruth Colwill and Judith Schrier for their valuable comments on an earlier draft of this manuscript. The work presented in this manuscript includes

studies from Melissa Acevedo's doctoral thesis, completed at Brown University in February 2002.

Correspondence about this article should be addressed to Joachim I. Krueger, Department of Psychology, Brown University, P.O. Box 1853, Providence, RI 02912 (e-mail: Joachim@Brown.edu). Received for publication July 30, 2004; revision received December 20, 2004.

1. Nozick (2001) reports that he stumbled on the assumption of common ground and worked out its implications in his doctoral dissertation at Princeton University in 1963.

2. This need also follows from the observation that mutual defection is the only Pareto-inefficient state of the game (i.e., there is a possible shift—to mutual cooperation—that would leave no one worse off).

3. For a given attribute that is either present or absent, the probability of two people being the same is $p^2 + (1 - p)^2$, where p is the probability of the attribute being present. If all players believe themselves to be in the majority, the percentage of correct predictions is at its maximum, which is p.

4. Setting w to .5 is generous. Van Lange's (1999) empirical estimates of w ranged from .13 to .32.

5. Bertrand Russell (1959) noted that the Chicken Game is an effective model of an arms race, in which mutual defection results in collective destruction. More prosaically, the same game is played by people who share (or try to get out of) responsibility to run a household or author a paper.

References

Acevedo, M., & Krueger, J. I. (2004). Two egocentric sources of the decision to vote: The voter's illusion and the belief in personal relevance. *Political Psychology, 25*, 115–134.

Andreoni, J. (1995). Cooperation in public-goods experiments: Kindness or confusion? *American Economic Review, 85*, 891–904.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*, 1390–1396.

Baker, F., & Rachlin, H. (2001). Probability of reciprocation in repeated Prisoner's Dilemma games. *Journal of Behavioral Decision Making, 14*, 51–67.

Baker, F., & Rachlin, H. (2002). Teaching and learning in a probability Prisoner's Dilemma. *Behavioral Processes, 57*, 211–226.

Brosnan, S. F., & de Waal, F. B. M. (2003). Monkeys reject unequal pay. *Nature, 425*, 297–299.

Clement, R. W., & Krueger, J. (2002). Social categorization moderates social projection. *Journal of Experimental Social Psychology, 38*, 219–231.

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

Dawes, R. M. (1989). Statistical criterion for establishing a truly false consensus effect. *Journal of Experimental Social Psychology, 25*, 1–17.

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.

Dawes, R. M., & Messick, D. M. (2000). Social dilemmas. *International Journal of Psychology, 35*, 111–116.

Dawes, R. M., & Orbell, J. M. (1995). The benefit of optional play in anonymous one-shot Prisoner's Dilemma games. In K. J. Arrow, R. H. Mnookin, L. Ross, A. Tversky, & R. Wilson (Eds.), *Barriers to conflict resolution* (pp. 63–85). New York: Norton.

Dawes, R. M., & Thaler, R. H. (1988). Anomalies. *Journal of Economic Perspective, 2*, 187–197.

De Cremer, D., & Van Lange, P. A. M. (2001). Why prosocials exhibit greater cooperation than proselfs: The role of social responsibility and reciprocity. *European Journal of Personality, 15*, S5–S18.

Dion, K. L. (1973). Cohesiveness as a determinant of ingroup–outgroup bias. *Journal of Personality and Social Psychology, 28*, 163–171.

Fantino, E., & Stolarz-Fantino, S. (2003). Experience and decision. *Behavioral and Brain Sciences, 26*, 160.

Gibbons, R. (1992). *Game theory for applied economists*. Princeton, NJ: Princeton University Press.

Goldman, A. I. (2001). Desire, intention, and the simulation theory. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 207–224). Cambridge, MA: MIT Press.

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

Hauert, C., & Doebell, M. (2004). Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature, 428*, 643–646.

Haxby, J., Parasuraman, R., Lalonde, F., & Abbound, H. (1993). Superlab: General purpose Macintosh software for human experimental psychology and psychology testing. *Behavior, Research Methods, Instruments, and Computers, 25*, 400–405.

Karniol, R. (2003). Egocentrism versus protocentrism: The status of self in social prediction. *Psychological Review, 110*, 564–580.

Kay, A. C., & Ross, L. (2003). The perceptual push: The interplay between implicit cues and explicit situational construals on behavioral intentions in the Prisoner's Dilemma. *Journal of Experimental Social Psychology, 39*, 634–643.

Kelley, H. H., & Stahelski, A. J. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology, 16*, 66–91.

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

Kiyonari, T., Tanida, S., & Yamagishi, T. (2000). Social exchange and reciprocity: Confusion or heuristic? *Evolution and Human Behavior, 21*, 411–427.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Reviews of Psychology, 24*, 183–214.

Komorita, S. S. (1976). A model of the *n*-person dilemma-type game. *Journal of Experimental Social Psychology, 12*, 357–373.

Komorita, S. S., & Parks, C. D. (1995). Interpersonal relations: Mixed motive interaction. *Annual Review of Psychology, 46*, 183–207.

Komorita, S. S., Parks, C. D., & Hulbert, L. G. (1992). Reciprocity and the induction of cooperation in social dilemmas. *Journal of Personality and Social Psychology, 62,* 607–617.

Komorita, S. S., Sweeney, J., & Kravitz, D. A. (1980). Cooperative choice in the *n*-person dilemma situation. *Journal of Personality and Social Psychology, 38,* 504–516.

Krueger, J. (1998). On the perception of social consensus. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 163–240). San Diego, CA: Academic Press.

Krueger, J. (2000a). Distributive judgments under uncertainty: Paccioli's game revisited. *Journal of Experimental Psychology: General, 129,* 546–558.

Krueger, J. (2000b). The projective perception of the social world: A building block of social comparison processes. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 323–351). New York: Kluwer.

Krueger, J. I., & Acevedo, M. (2005a). *Perceptions of self and other in the Prisoner's Dilemma game.* Unpublished manuscript, Brown University.

Krueger, J. I., & Acevedo, M. (2005b). 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., & DiDonato, T. (2005). *Intentions to vote in the 2004 presidential election.* Unpublished manuscript, Brown University.

Kuhlman, D. M., & Wimberley, D. L. (1976). Expectations of choice behavior held by prosocials, competitors, and individualists across four classes of experimental game. *Journal of Personality and Social Psychology, 32,* 922–931.

Laplace, P. S. (1814). *Essai philosophique sur les probabilities* [Philosophical essay on probabilities]. Paris: Courcier.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Luce, R. D., & Raiffa, H. (1957). *Games and decisions.* New York: Wiley.

Messé, L. A., & Sivacek, J. M. (1979). Predictions of others' responses in a mixed-motive game: Self-justification or false consensus? *Journal of Personality and Social Psychology, 37,* 602–607.

Messick, D. M., Bloom, S., Boldizar, J. P., & Samuelson, C. D. (1985). Why we are fairer than others. *Journal of Experimental Social Psychology, 21,* 480–500.

Messick, D. M., & McClintock, C. G. (1968). Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology, 4,* 1–25.

Monterosso, J., & Ainslee, G. (2003). Game theory need not abandon individual maximization. *Behavioral and Brain Sciences, 26,* 171.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Nozick, R. (1993). *The nature of rationality.* Princeton, NJ: Princeton University Press.

Nozick, R. (2001). *Invariances: The structure of the objective world.* Cambridge, MA: Harvard University Press.

Onorato, R. S., & Turner, J. C. (2004). Fluidity of the self-concept: The shift from personal to social identity. *European Journal of Social Psychology, 34,* 257–278.

Orbell, J., & Dawes, R. M. (1991). A "cognitive miser" theory of cooperators' advantage. *American Political Science Review, 85,* 515–528.

Orbell, J. M., & Dawes, R. M. (1993). Social welfare, cooperators' advantage, and the option of not playing the game. *American Sociological Review, 58,* 787–800.

Parks, C. D., Sanna, L. J., & Berel, S. R. (2001). Actions of similar others as inducements to cooperate in social dilemmas. *Personality and Social Psychology Bulletin, 27,* 345–354.

Quattrone, G. A., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology, 46,* 237–248.

Rachlin, H. (2002). Altruism and selfishness. *Behavioral and Brain Sciences, 25,* 239–291.

Rapoport, A. (1967). A note on the index of cooperation for Prisoner's Dilemma. *Journal of Conflict Resolution, 11,* 101–103.

Rapoport, A. (1987). Research paradigms and expected utility models for the provision of step-level public goods. *Psychological Review, 94,* 74–83.

Rapoport, A. (2003). Chance, utility, rationality, equilibrium. *Behavioral and Brain Sciences, 26,* 172–173.

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.

Russell, B. (1959). *Common sense and nuclear warfare.* New York: Simon & Schuster.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958–1992. *Rationality and Society, 7,* 58–92.

Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.

Schelling, T. C. (1960). *The strategy of conflict.* Cambridge, MA: Harvard University Press.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology, 24,* 449–474.

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. (2000). Beyond self-interest: A set of propositions relevant to interpersonal orientations. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 11, pp. 279–331). Chichester, UK: Wiley.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly, 63,* 116–132.