Journal of Conflict Resolution
Volume 51 Number 3
June 2007 431-456

10.1177/0022002707300834
http://jcr.sagepub.com
hosted at
http://online.sagepub.com

Speaking the Same Language

The Cooperative Effects of Labeling in the Prisoner's Dilemma

Chen-Bo Zhong
Rotman School of Management
University of Toronto, Canada

Jeffrey Loewenstein
McCombs School of Business
University of Texas, Austin

J. Keith Murnighan
Kellogg School of Management
Northwestern University, Evanston, Illinois

The long history of experimental research on the prisoner's dilemma (PD) has primarily used a methodology that eliminates cues to participants. Researchers, however, have interpreted participants' choices as cooperative or competitive. The authors' research shows that giving participants researchers' interpretive labels of the game, the choices, and the outcomes, compared to no labels, led to significantly more cooperation; labels such as *trust* and *cooperate/defect* augmented cooperation even more. A second experiment found that independent evaluations of the labels led to perceptions that were similar to individuals' choices in the first experiment. These results suggest that we might need to rethink the import of many of our previous findings and their applicability to everyday interactions.

Keywords: *prisoner's dilemma game; cooperation; trust; labeling*

Since A. W. Tucker (1950) introduced the prisoner's dilemma (PD) game, scholars in a wide array of fields (psychology, economics, anthropology, political science, and others) have investigated its dynamics. Analogs of the experimental PD are numerous, ranging from price wars in business to evolutionarily stable symbiosis (Ostrom et al. 2002). Although interest in this game has ebbed and flowed, often in the face of withering criticism (e.g., Nemeth 1972), researchers continue to use PD games as the basis for their research (e.g., Bottom et al. 2002).

Authors' Note: We are particularly thankful for the generous research funding provided by the Dispute Resolution Research Center (DRRC) in the Kellogg School of Management at Northwestern University. Data for replication are available at http://jcr.sagepub.com/cgi/content/full/51/3/431/DC1/, alongside the electronic version of this article.

The essential structure of the PD depends on the game's payoffs: for any one play of the game, cooperative choices increase everyone's payoffs, but noncooperative choices increase noncooperators' individual payoffs. Formal models (e.g., Luce and Raiffa 1957) suggest that finite PD games should not elicit much cooperation. Casual observations of naturalistic interactions that conform to the definitional elements of the PD, however, suggest that people often cooperate. Thus, charities and churches receive many contributions, work groups often achieve their goals, and voluntary constraints actually reduce consumption. The inconsistencies between life in our theories and our laboratories and life among our peers are at once troubling and hopeful.

The effects of juxtaposition and the interrelationships of the payoffs in the PD matrix has been a central concern in many research studies. To carefully isolate these effects, researchers have followed the basic, elemental philosophy of experimental methods (e.g., Aronson, Wilson, and Brewer 1998). In other words, PD researchers (with some notable exceptions, e.g., Orwant and Orwant 1970) have typically removed informational cues that might provide richer meaning to the game. The absence of cues extends to (1) the games themselves, which are neither described in any detail nor named; (2) the players' choices, which receive only abstract labels (e.g., A and B); and (3) their outcomes, which are only expressed in payoff units (points, cents, etc.). The result is an emphasis on the payoff numbers, free of almost all contextual signals. Under these sparse conditions, Dawes's (1980) and Rapoport's (1988) reviews of PD and social dilemma research note that a baseline expectation for cooperation rates among anonymous strangers should be moderate—around 50 percent—with considerable variation. Research has also consistently shown that, in iterated games, players choose less cooperatively as the game's end approaches, especially on the last trial (e.g., Ledyard and Palfrey 1995).

The intent of this cueless methodological approach is to remove extraneous effects and provide a clean testing ground for models and hypotheses. Researchers taking this approach generally acknowledge potential external validity concerns but pay little attention to a second potential concern: participants' actual interpretations of the tasks. Despite Kelley and Thibaut's (1978) conceptual analysis of matrix games, which suggested that experimental participants take the given matrix and *transform* it before making their choices, PD research to date has rarely considered how participants have transformed the game. Instead, when we and other researchers have interpreted the players' choices, we have imposed our own interpretations on them. Although the players' choices might only reflect simple, payoff-relevant selections, we have interpreted players' choices as cooperation, noncooperation, defection, or competition. Outcome labels exhibit little variation as well: in their seminal early work, Rapoport and Chammah (1966) referred to the outcomes in the PD game as R (for reward), P (for punishment), T (for temptation to defect), and S (for the sucker's payoff). Subsequent usage (e.g., Murnighan and Roth 1983) referred to T as the traitor's payoff and S as the sucker or saint's payoff (depending on one's point of view). In other words, while our research designs

have steadfastly removed all labels from participants' view, we consistently used these labels to interpret their behavior. As Rapoport (1988, 464) put it, "These interpretations are standard." Thus, research on the PD may have suffered from a clear disconnect between the way that the games have been presented to the players and the way that researchers have interpreted their choices and outcomes. In essence, the games that we felt that we were studying may have been systematically different from the games that our participants thought that they were playing.

Thus, PD researchers and participants may not have been thinking or speaking the same language as they have interpreted the games, the choices, and the outcomes. The purpose of the current experiments was to compare and contrast the effects of different descriptions of the essential elements in the structures of PD games, that is, the game itself, the players' choices, and their outcomes. In particular, these experiments investigated whether providing participants with the same labels that researchers have used to interpret their play might change their outlook in PD games, as expressed in their evaluations of the game and their behavioral choices. By allowing participants to speak the same language that researchers have been using, these studies attempted to minimize transformational differences that might have biased previous interpretations.

The broader implications of this work are twofold. First, there are likely to be theoretical and practical benefits in refining or injecting greater variety in PD vocabularies. Alison, Beggan, and Midgely (1996), for instance, argued for the benefits of multiple metaphors for social dilemmas, suggesting that the breadth and fruitfulness of research hypotheses was driven in part by the contextual variety in which dilemmas had been examined. Second, there is a need for greater attention on how people interpret decision situations (Tor and Bazerman 2004; see also research on framing effects, e.g., Tversky and Kahneman 1986). Language routinely influences how people understand their decisions: "understanding decision making involves understanding the ways in which language carries, elaborates, and creates meaning" (March 1994, 211-12).

Previous Research

Researchers have been remarkably consistent in their attempts to remove any and all cues from the choices or the outcomes in their experiments. To support this contention in an enormous literature on PD games (van Lange et al. 1992), we chose three representative reviews to document how researchers formulated their experiments and interpreted their participants' choices. We used one review from the 1970s (Wrightsman, O'Connor, and Baker 1972), one from the 1980s (Messick and Brewer 1983), and one from the 1990s (Komorita and Parks 1995), as well as our own previous research. These sources cited forty experiments that included some *n*-person PD games but primarily two-person games. Reviewing the original articles indicated that participants' choices were variously labeled A and B; C and

D; X and Y; red, blue, yellow, or white; up and down; or left and right. Two of our studies (Bettenhausen and Murnighan 1991; Murnighan, King, and Schoumaker 1990) departed from the norm, presenting experimental participants with PD games described as bidding games that required high or low bids. The conclusion from this review is simple: the vast majority of research presented abstract, minimal labels for choices, and none for outcomes or the PD game itself.

A small set of PD experiments have provided participants with labels for the entire game. The central finding is that people cooperated more in PD games that were (1) labeled a "Community Game" rather than a "Wall Street Game" (Liberman, Samuels, and Ross 2004), (2) described as an international negotiation rather than an economic bargaining situation (Eiser and Bhavnani 1974), and (3) labeled as a "Social Exchange Study" rather than a "Business Transaction Study" (Batson and Moran 1999). Although these studies suggest a clear pattern of results, none of these studies included unlabeled baselines to ascertain whether the labels provided led to cooperation increases or declines, or both. We also found three PD experiments that labeled participants' choices explicitly as *cooperate* and *defect* or *compete* (Johnston, Markey, and Messe 1973; Oskamp and Perlman 1965; Rilling et al. 2002); none of these studies yielded strong evidence of behavioral differences. Finally, we found no studies that labeled participants' outcomes, that is, none that identified the outcomes as rewards, punishments, or payoffs for saints, suckers, or traitors. In short, although researchers have a well-articulated vocabulary for describing PD games, choices, and outcomes, they have only rarely shared it with their participants.

Research on social dilemmas provides results that are consistent with the data from the three studies that labeled PD games. For example, people contributed more in a multitrial social dilemma game if it was a "Social Event Fund Decision" rather than a "Joint Investment Fund Decision" (Pillutla and Chen 1999). Labeling social dilemma choices has also influenced participants' behavior. The opportunity to *give*, *take*, *save*, *contribute*, or *harvest* their resources or outcomes, for instance, has led to different effects in structurally equivalent games. "Take-some" games, for instance, generate more cooperation than mathematically identical "give-some" games (Dawes 1980): people appear less willing to accept a loss than to forgo a formally equivalent gain (consistent with prospect theory; Kahneman and Tversky 1979).

Larrick and Blount's (1997) research further clarified these points. They noted that choices in social dilemma and ultimatum games were theoretically identical. (In an ultimatum game, one party makes an offer to another from a total payoff; if the second party accepts, they split the total as specified by the offerer; if the second party rejects, both parties receive nothing.) In a pointed series of experiments, they found that the act of *accepting and rejecting* offers in ultimatum games raised concerns for fairness, control, and responsibility more than did the act of *claiming* in social dilemmas, and these different conceptualizations accounted for the marked differences in cooperation in these two otherwise identical games.

Research on labels and on the construal of games, choices, and outcomes, however, is not common. Thus, although participants' choices in a recent social dilemma study (van Dijk and Wilke 2000) were described as *give*, *keep*, *leave*, or *take*, summary results for the four labels were not reported; instead, they were collapsed into theoretically imposed categories. At the same time, dilemma researchers (e.g., Ostrom et al. 2002) are actively examining perceptions that mediate cooperation, such as trust or fairness. This suggests the need to understand how those perceptions are triggered in the first place. We suggest that the labels that people are given for their interactions, their choices, and their outcomes are likely to have important effects on their interpretations and choices in PD games and decision making more broadly.

Experiment 1

Our underlying question was whether the labels that researchers (ourselves included) have used to describe PD games accurately characterize participants' interpretations of these games, as displayed in their game choices. Accordingly, this experiment provided clear and strongly worded labels for the PD game, for the players' choices, and for their outcomes: blunt labels—*trust* and *cutthroat*—described the game in starkly different ways; the choice labels reflected researchers' common interpretative descriptions of participants' choices, for example, *cooperate*, *defect*, *rational*, and *choose for the group*; and labels for the participants' outcomes also reflected common research usage, for example, *winner*, *sucker*, *saint*, *punishment*, or *group maximum*. Conditions with no labels were also included to assess whether labels promote or diminish cooperation relative to the conditions in typical PD experiments. Participants played six 12-trial games to assess the effects of game, choice, and outcome labels and to determine whether these labels had a consistent influence across trials.

Schelling (1960) argued that rational players can and should use available labels as cues for predicting and understanding people's actual and potential choices. If the researchers' interpretive labels accurately portray participants' construals of the game, their choices, and their outcomes, there should be no difference in the play of participants who do and do not receive labels. Alternatively, if participants ignore the labels and only consider the numbers in the payoff matrices, there should also be no effects for labels. If, however, researchers' and participants' interpretations of these interactions differ, then participants' choices should clearly be affected when they play PD games using researchers' interpretive labels. Accordingly, we expected that positive, group-oriented labels would increase cooperation and that negative, self-oriented labels would decrease cooperation relative to cooperation in unlabeled games.

This was most easily investigated for game labels, that is, trust and cutthroat. The labels that researchers have used to describe the players' choices and their

Table 1
Mean Levels of Cooperation for the Different Labels across Conditions (in percentages)

	All Games	Game 1 (n = One-Sixth of the Size for All Games)	Cooperation Decline in the Final Two Trials of Each Game (Endgame), All Games
Game labels (n = 288)			
Trust	63**	48	15
Cutthroat	58	52	16
Game Blank	59	55	13
Choice labels (n = 144)			
Cooperate/Defect	64**	59	14
Cooperate/Don't Cooperate	64**	49	13
Don't Defect/Defect	60	43	15
Choose for Group/Self	58	51	13
Idealistic/Rational	59	52	19
Choice Blank	58	53	13
Outcome labels (n = 216)			
Saint	64	60**	7**
Sucker	66	58**	16
Maxmin	54	45	16
Outcome Blank	58	42	20
Total (n = 1,002)			
Labels (mean of all) (n = 864)	60**	51	15
Baseline (unlabeled) (n = 138)	48	44	12

Note: The Game Blank, Choice Blank, and Outcome Blank conditions were not completely blank. For instance, the Game Blank condition included games with choice and outcome labels. The Baseline (unlabeled) condition included no labels of any kind.

**Significantly different from the comparable blank label group, $p < .05$.

outcomes also have positive and negative connotations, but the fact that there are two choices and four outcomes means that the meanings that can be ascribed to choices and outcomes are more complex. Table 1, for instance, displays the choice labels that we used in this experiment. Several pairs of choice labels provide positive connotations for the first choice (cooperate, don't defect) and negative connotations for the second choice (defect and don't cooperate).

Outcome labels are the most complex if only because they include four different labels. The combination that we have labeled Saint (including winner, traitor, saint, and punishment) includes two positive and two negative labels, all of which encourage cooperation and discourage defection. The combination that we have labeled Sucker (including winner, traitor, sucker, and punishment) includes only one positive. Thus, it might stimulate less cooperation than the Saint labels. The third combination of group and individual maxima and minima also includes two positive

and two negative labels, but the positives (the maxima) or the negatives (minima) result as a function of the other's choice. Multiple labels also offer participants the opportunity to focus more on some labels and less on others, allowing individuals to frame their own choices positively, even if their counterpart might not view them so positively.

A final question for this research was whether the introduction of labels would qualitatively shift participants' pattern of choices. Thus, in addition to overall rates of cooperation, we focused on the typical tendency of participants to cooperate less frequently in the endgame (Ledyard and Palfrey 1995). Finding a reduction or an acceleration of the endgame drop in cooperation with different labels would further suggest that participants transform the payoffs, incorporating their own contextual meanings in ways that differ from researchers' typical conceptualizations.

Method

Participants

The participants were 167 Northwestern University undergraduate student volunteers who responded to notices around campus. About half (83; 49.4 percent) were males; 47 percent (79) were females; 5 participants did not provide gender information.

Design

Each participant played six different twelve-trial games that varied the game, choice, and outcome labels. Two of the games were labeled Trust, two Cutthroat, and two were unlabeled (blank). The player's two choices either had no labels (blank) or one of five other combinations: Cooperate-Defect; Cooperate–Don't Cooperate; Idealistic-Rational; Don't Defect–Defect; and Choose for Group–Choose for Self. Participants experienced each of the six sets of choice labels in one of their six games. Participants' outcomes had either no labels or one of three sets of labels: Winner-Saint-Traitor-Punishment, Winner-Sucker-Traitor-Punishment, and Group Maximum–Individual Minimum–Individual Maximum–Group Minimum (see Table 1). All participants saw only one set of outcome labels in all of their six games.

Because the three independent variables had many levels, we used a mixed design: outcome labels (four levels) were a between-subject factor; game (three levels) and choice labels (six levels) were within-subject factors. Completely crossing the within factors would have required each participant to experience eighteen games. Instead, each participant played six games, experiencing each of the choice labels once and each of the game labels twice. To control for order effects, we generated two Latin Squares, a 6×6 Latin Square for the order of the choice labels and a 3×3 Latin Square for the order of the game labels. Cross-multiplying these

two Latin Squares yielded eighteen different conditions. Factorially crossing these with the outcome factor (four levels) led to seventy-two different conditions. Two participants were randomly assigned to each of these seventy-two combinations condition, for a total of 144 of our experimental participants. The incomplete mixed design and incomplete Latin Square does not allow for a test of the two-way interaction between game and choice labels or the three-way interaction between game, choice, and outcome labels. However, tests for main effects of the three variables and the other interactions (i.e., outcome and game label, outcome and choice label) are unbiased, and these were the main foci of this study.

Only one of the seventy-two conditions included no labels for games, choices, and outcomes. Because a completely unlabeled game is an important standard for comparison and because it represents the format of most previous PD research, we added a baseline condition in which twenty-three participants played six games with no labels of any kind.

The payoffs in each of the six games were chosen to be relatively neutral in terms of motivations to cooperate or defect. With respect to various indexes that have been suggested to identify the cooperativeness of a game (Murnighan and Roth 1983), the payoffs used here are in the midrange for r_1 and r_2 (Rapoport and Chammah 1966), r_3 and r_4 (Harris 1969), and e_1 and e_2 (Roth and Murnighan 1978). Six sets of payoffs were constructed; they differ from each other due to the addition or subtraction of a constant. Thus, they are monotonic transformations of one another, that is, theoretically identical (see Table 2). The six sets of payoffs were randomly paired with the label conditions. Each participant experienced each of them once.

Participants' mean number of cooperative choices for each game was our main dependent variable. We also contrasted the rate of cooperation across the first ten trials of each game (the ongoing rate of cooperation) with that on the last two trials of each game (the endgame rate of cooperation). Supplementary analyses related participants' performance to their responses on a postexperiment questionnaire (shown in the appendix).

Procedure

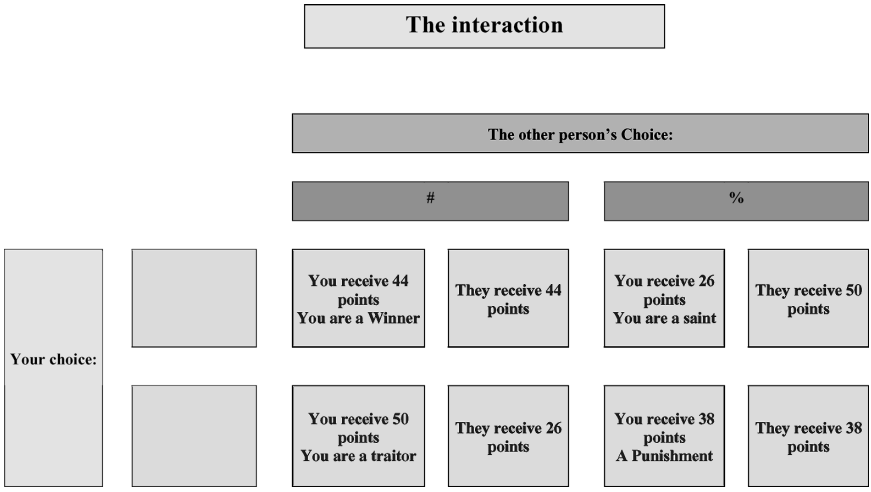
Upon arrival at the lab, each participant was led to either a computer-equipped cubicle or a small room. Participants were led to believe that they were interacting with each other via computer, ostensibly to preserve anonymity. Actually they interacted with a programmed set of choices.

Participants played six games, each consisting of twelve repeated trials, with six different "individuals." Participants were encouraged to maximize their outcomes; they received outcome feedback after each trial. They were told that their monetary payoff for participation would be determined by their performance in one of the six games and that it would be randomly selected. Thus, they were encouraged to do well for themselves in each and every game.

Table 2
Payoff Structures and Choice and Outcome Labels (in parentheses) for the Six Games

		Other's Choice			Other's Choice	
		Cooperate	Defect		Cooperate	Don't Cooperate
Your Choice	Cooperate	44, 44 (winner)	26, 50 (saint)	Cooperate	30, 30 (winner)	12, 36 (sucker)
	Defect	50, 26 (traitor)	38, 38 (punishment)	Don't Cooperate	36, 12 (traitor)	24, 24 (punishment)
		Choose for the Group	Choose for the Self		Idealistic	Rational
Your Choice	Choose for the Group	27, 27 (group maximum)	9, 33 (individual minimum)	Idealistic	36, 36	18, 42
	Choose for the Self	33, 9 (individual maximum)	21, 21 (group minimum)	Rational	42, 18	30, 30
		Don't Defect	Defect		@	#
Your Choice	Don't Defect	24, 24	6, 30	@	40, 40	22, 46
	Defect	30, 6	18, 18	#	46, 22	34, 34

Figure 1
A Reproduction of the Instructional Screen, Experiment 1



Before the games began, participants saw an instruction screen that contained a typical PD game (see Figure 1) and read an explanation of the mechanics of the experiment (i.e., how to use the mouse to make their choices). The labels in the computer screen's boxes manipulated the independent variables during the experiment. For the practice trials used to familiarize participants with the game, the instructions used neutral symbols such as "#" and "*" for their choices and "Interaction #1" for the game. The outcome labels that were used in their condition of the experiment were used in their instructions. Finally, the payoffs in the practice trials conformed to the requirements of a PD game but differed from those used in the actual experiment. Participants were told to take enough time to read the instructions and ask any questions before proceeding.

Participants were then told that their first game would be either a Trust Game or a Cutthroat Game, or they were given no description of the game. In the label conditions, participants saw computer screens that highlighted their choices and outcomes in boxes of different colors. Participants then received a message telling them to get ready while the computer randomly matched them with their counterpart. Their computer screen looked just like the one they had seen in the instructions, but with the actual payoffs displayed for that particular game and the labels that were appropriate for the experimental conditions.

In the baseline, no-label condition, both choice and feedback label boxes were empty, and the game label box was called a numbered interaction (e.g., Interaction

#3). In the label conditions, the box at the top of the screen was filled with the words "Trust game" or "Cutthroat game," or a numbered interaction. Similarly, the boxes next to "your choice" held labels for the players' choices, such as "Cooperate" and "Defect," or merely blank boxes. Finally, outcome labels were included in the boxes that displayed their outcomes in points. Thus, they might have seen "You received 50 points. You are a traitor."

Participants did not have time limits. We found that they tended to take longer on the first trial of each game (about eight seconds) and less on subsequent trials (down to about two seconds). Following each of their choices, participants saw a message that the computer was recording both players' choices. This message lasted from one-half to three seconds and was programmed to vary across trials within each game. Participants then received their outcome for that trial, with the appropriate label. They made choices with the same payoffs with the same "counterpart" for the whole game of twelve trials. They were then given a summary of the points they had accumulated in that game prior to proceeding to their next game with, supposedly, a different counterpart. Play across the six consecutive games provided additional information on whether participants learned to be more or less cooperative.

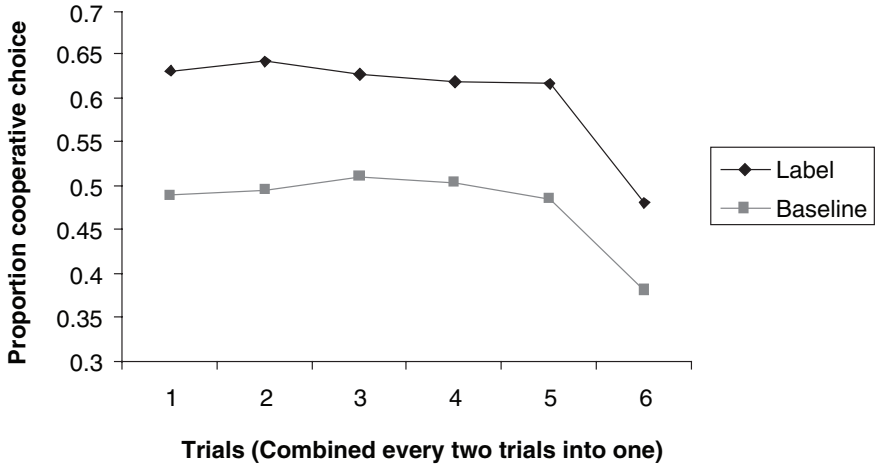
In all of the games, the computer program chose cooperatively on the first trial of the game. For the next nine trials of the game, it used the tit-for-tat strategy, taking the participant's choice on the previous trial as its choice for the current trial. On the last two trials of the game, the program matched the participants' current choice rather than echoing their choice[1] on the previous trial.

After the last game, participants were asked to complete a short questionnaire (see the appendix). As manipulation checks, participants were asked to recall game, choice, and feedback labels. Open-ended questions asked them to describe the focus of their attention and their strategies in the experiment. Scaled questions collected demographic information and whether they were familiar with the PD game, whether they had had any game theory classes, and how many economics classes they had taken. After completing the questionnaire, participants saw a screen containing their outcomes in all six games. The experimenter then entered the room with six lottery tickets and let participants randomly choose a number from one to six to determine their "payoff" game. Participants' pay ranged from $14 to $16 for approximately forty-five minutes. We invited participants' questions in a short debriefing session. None of the participants indicated any knowledge of the hypotheses.

Results

Adding labels increased participants' cooperative choices: participants who saw labels for their game, their choices, and/or their outcomes cooperated more than participants in the no-label, baseline condition: 60 versus 48 percent cooperation, $F(1, 1{,}000) = 14.78$, $p < .001$ (see Figure 2 and Table 1).[2]

Figure 2
The Mean Proportion of Cooperative Choices by Pairs of Trials for Labeled and Unlabeled Prisoner's Dilemma Games, Experiment 1



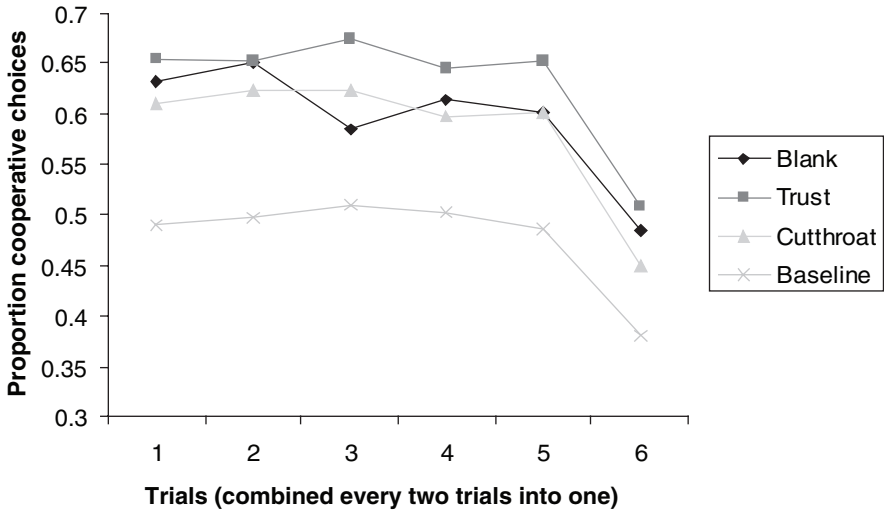
Game Labels

Each of the game labels led to significantly more cooperation than the baseline, unlabeled condition (HSD[3] max $p < .05$; see Figure 3 and Table 1). The three game labels themselves differed in the rates of cooperation they engendered, $F(2, 280) = 4.69$, $p < .01$. As expected, participants cooperated more in the trust game than the cutthroat and blank games, $F(1, 286) = 8.66$, $p < .05$; and $F(1, 286) = 5.24$, $p < .05$, respectively. Overall, the cooperation rate was comparable for the cutthroat and blank games, $F(1, 286) = .42$, $p > .10$.[4] Finally, choices of all of the game label participants showed comparable declines of approximately 15 percent cooperation on the last two trials.

Choice Labels

Each of the six choice labels led to significantly more cooperation than the baseline, unlabeled condition (Tukey HSD max $p < .05$; see Figure 4 and Table 1). There were marginally significant differences across the different choice labels, $F(5, 715) = 2.11$, $p = .062$. Contrast tests indicate that including the word *cooperate* (cooperate/defect: 64 percent; and cooperate/don't cooperate: 64 percent) led to more cooperative choices than the blank choice condition: 58 percent, $F(1, 715) = 5.11$, $p < .05$; and $F(1, 715) = 5.82$, $p < .05$, respectively. The remaining choice labels (don't defect/defect: 60 percent; idealistic/rational: 59 percent;

Figure 3
The Mean Proportion of Cooperative Choices by Pairs of Trials for the Game Labels, Experiment 1

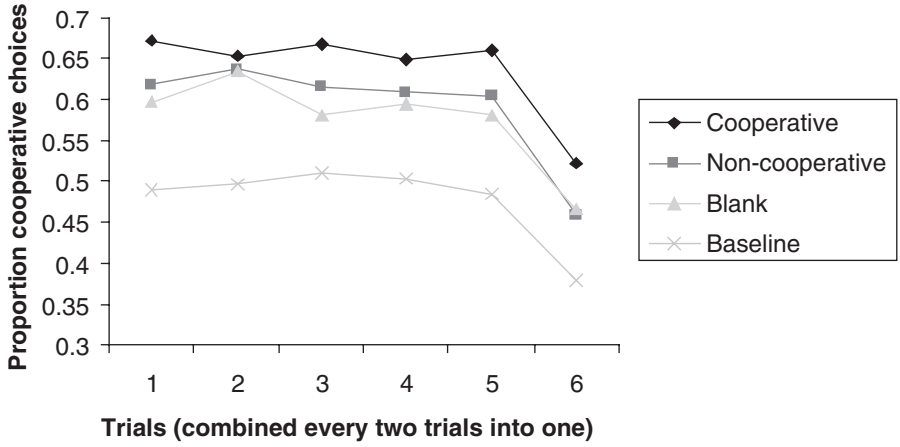


group/self: 58 percent) did not differ significantly from the blank choice condition or one another. All choice labels led to about 14 percent drops in cooperation on the last two trials except the idealistic/rational labels, which resulted in a larger (19 percent) but not significantly larger drop.

Outcome Labels

Outcome labels were varied between subjects instead of within subjects, which meant that the analyses had less power. Overall, the outcome labels led to marginally more cooperation than the baseline condition, $F(1, 162) = 3.75$, $p = .054$.[5] Despite a trend for greater cooperation in the saint (64 percent) and sucker (66 percent) conditions than in the maxmin (54 percent) and blank (58 percent) conditions, the outcome labels did not differ significantly from each other, $F = 1.18$, n.s. In the first of the six games, the saint (60 percent) and sucker (58 percent) labels both led to more cooperation than the blank outcome labels: 42 percent, $F(1, 140) = 5.03$, $p < .05$; and F (1, 140) = 4.27, $p < .05$, respectively (see Figure 5 and Table 1). The outcome labels also had significantly different effects on the last two trials, $F(15, 700) = 1.66$, $p = .05$. Cooperation in the saint condition fell significantly less than the drop in the other outcome label conditions (declining 7 percent versus an average of 17 percent; min $F > 6$, $p < .015$).

Figure 4
The Mean Proportion of Cooperative Choices by Pairs of Trials for the Choice Labels, Experiment 1



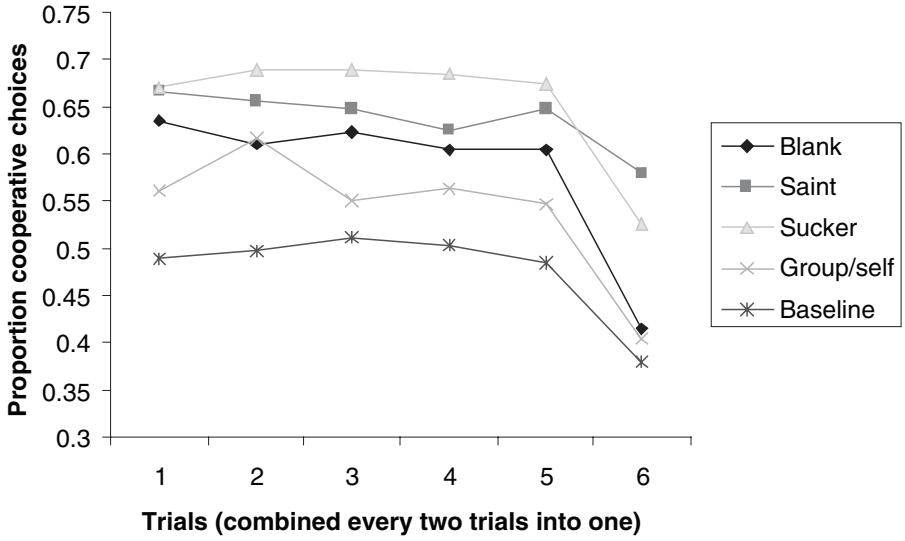
Learning

Over the course of the six games, participants seemed to learn that cooperative choices were beneficial. They did not, however, discover or learn that they could also cooperate on the last two trials of each game (even though their programmed opponent never defected on these trials unless they did). Participants' mean level of cooperation increased from 51 percent in the first game to 64 percent in their sixth game, a significant increase, $F(1, 715) = 44.56$, $p < .001$. A contrast between the first and sixth games was also significant, $F(1,715) = 23.99$, $p < .005$. Participants in their first game maintained most of their cooperation through the last two trials of the game (dropping just 7 percent). By the sixth game, cooperation on the last two trials of the game was diminishing much more (dropping 16 percent), leading to a significant trials by game order interaction, $F(25, 4,150) = 1.84$, $p < .01$. Interaction contrasts indicated that the end of game drop in cooperation in the first game was significantly different from all of the others (min $F > 5.6$, $p < .025$) and that the second and sixth games, $F(1, 4,150) = 3.93$, $p < .05$; and third and sixth games, $F(1, 4,150) = 3.85$, $p < .05$, also differed significantly (see Figure 6).

Postexperiment Questionnaire

Data from the postexperiment questionnaire suggested that participants differentially attended to the labels. Most participants reported that they primarily attended to

Figure 5
The Mean Proportion of Cooperative Choices by Trial for the Outcome Labels, Experiment 1

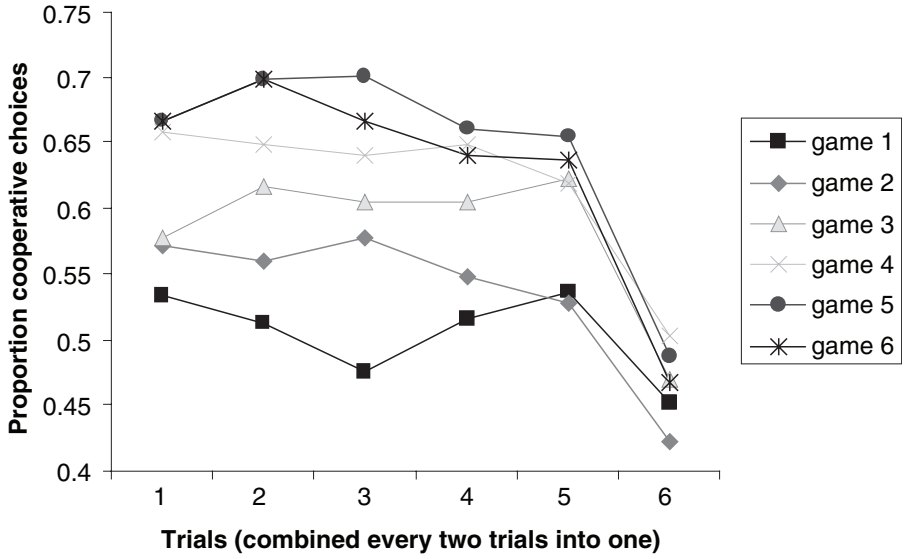


the numbers rather than to the words, with 46 percent rating themselves a 7 on a scale from 1 = *primarily words* to 7 = *primarily numbers*. Those rating themselves less than 6 ($n = 49$) cooperated more (67 percent) than those rating themselves a 6 or 7 ($n = 89$; 56 percent), $t(136) = 2.06$, $p < .05$. Unlike findings from previous research, which indicated that more economics classes led people to cooperate less in the PD game (Frank, Gilovich, and Regan 1993), participants who were familiar with game theory and PD games cooperated more than those who were not (64 vs. 54 percent; $t = 2.15$, $p < .05$). Familiarity with game theory and PD games, however, did not seem to affect the influence of labels. We divided the sample into low and high familiarity with PD games[6] and tested whether this dichotomous variable influenced any of the effects of the labels. We found that familiarity with PD games did not interact with any of the game, $F(2, 282) = .90$, $p = .41$; choice, $F(5, 703) = 1.31$, $p = .26$; or outcome labels, $F(3, 158) = 1.0$, $p = .39$. This suggests that labels influenced participants' choices regardless of whether they were familiar with PD games.

Discussion

We used a large, omnibus design in this experiment to determine if labeling the game, the choices, or the outcomes might lead to different results from the typical,

Figure 6
The Mean Proportion of Cooperative Choices by Pairs of Trials for the Six Games in the Order of Their Play, Experiment 1



unlabeled conditions of previous PD research. Clearly, the findings indicate that they do.

For many years, while researchers have avoided using labels in their PD experiments, they have used standard labels to interpret their participants' choices. The current findings suggest that participants interpreted the games differently: their choices may not have been consistent with our interpretations of their choices. When this experiment provided participants with the labels that we have traditionally used to interpret their behavior, they were significantly more cooperative (although we now know that we must use the term *cooperative* cautiously). Clearly, our transformations of the payoff matrices have not been consistent with our participants' transformations.

Even though cooperation in PD games represents a risky choice, participants in this study responded positively and immediately to the implicit encouragement of choice labels like cooperate/defect. Thus, in the very first trial of their first game, participants who played a Trust game cooperated 60 percent of the time ($n = 48$), participants who had Cooperation as one of their two choices cooperated 71 percent of the time ($n = 24$), and participants who had one of their four outcomes labeled Saint cooperated 72 percent of the time ($n = 36$). Although these figures represent

relatively small samples, and cooperation rates did not remain this high for long (dropping to 48, 54, and 60 percent, respectively, in their first games), in all three cases, cooperation was considerably greater than it was in the first trial of the baseline, unlabeled game (39 percent).

In most games participants saw multiple labels at once. There are a few games in which only one label varied. Thus, participants who played a "Trust game" with blank choice and outcome labels cooperated 60 percent of the time ($n = 12$), compared to 63 percent cooperation when the label "Trust Game" was present (along with other labels) and 48 percent in the baseline game. Those who played a "Cutthroat game" with no other labels, however, cooperated only 35 percent of the time. Although this represents only twelve games, it suggests that this label led to less cooperation than the baseline, unlabeled condition. Likewise, the impact of outcome labels was amplified if they were presented alone. Participants with the Saint (58.3 percent) or Sucker (75.7 percent) outcome labels only cooperated more than those with the maxmin outcome labels only (47.9 percent) or baseline (48 percent). Although limited sample sizes prevent conclusive analyses, these cases suggest that the effect of a particular label may be diluted by the presence of other labels. Thus, our results were considerably conservative (we return to this issue in the general discussion).

These labels influenced people even if they did not realize that they had been influenced, as most of our participants (64.5 percent) said that they focused entirely on the payoffs and ignored the labels. Their cooperation rates, however, were well above those in the baseline condition (56.2 vs. 47.8 percent). Furthermore, the Saint outcome label even appeared to mitigate the typical endgame drop in cooperation rates. It is possible that this intriguing discrepancy between perception and behavior is due to a social desirability bias—that participants wanted to be seen as rational number crunchers. It is also possible that labels can exert influence on behaviors without conscious attention, which makes the power of labels more impressive—if people do not realize that their behaviors are affected by labels, how could they possibly resist such influences?

The minority of participants (about a third) who did say that labels mattered tended to express this in their open-ended responses at the end of the study. They said they felt badly about being a "traitor" guilty of "defecting," and instead wanted to be a "winner," to "cooperate," or to achieve the "group maximum." Some suggested that maintaining positive labels was in tension with or was more important than obtaining payoffs. For instance, one participant wrote, "I like winning points mostly, but I also somewhat like to be called someone who cooperates." And another wrote, "I just always clicked the good cooperative button . . . the extra 6 points didn't mean enough to me."

Thus, labels can influence people's choices in many ways, both implicitly and explicitly. Given the current evidence, it seems clear that these labels gave people potent signals that suggested particular interpretations, which then resulted in