ELSEVIER

Evolution and Human Behavior 25 (2004) 63–87

Evolution and Human Behavior

Third-party punishment and social norms

Ernst Fehr*, Urs Fischbacher

Institute for Empirical Research in Economics, University of Zurich, Blümlisalpstrasse 10, CH-8006 Zürich, Switzerland

Received 2 July 2003; accepted 28 January 2004

Abstract

We examine the characteristics and relative strength of third-party sanctions in a series of experiments. We hypothesize that egalitarian distribution norms and cooperation norms apply in our experiments, and that third parties, whose economic payoff is unaffected by the norm violation, may be willing to enforce these norms although the enforcement is costly for them. Almost two-thirds of the third parties indeed punished the violation of the distribution norm and their punishment increased the more the norm was violated. Likewise, up to roughly 60% of the third parties punished violations of the cooperation norm. Thus, our results show that the notion of strong reciprocity extends to the sanctioning behavior of "unaffected" third parties. In addition, these experiments suggest that third-party punishment games are powerful tools for studying the characteristics and the content of social norms. Further experiments indicate that second parties, whose economic payoff is reduced by the norm violation, punish the violation much more strongly than do third parties.

Keywords: Social norm; Sanction; Punishment; Strong reciprocity; Social preference; Third party

1. Introduction

No human societies exist without social norms, that is, without normative standards of behavior that are enforced by informal social sanctions. In fact, the ability to develop and enforce social norms is probably one of the distinguishing characteristics of the human species. It is, therefore, not surprising that social scientists, at least those outside of

* Corresponding author. Tel.: +41-1-634-3709; fax: +41-1-634-4907.
E-mail addresses: efehr@iew.unizh.ch (E. Fehr), fiba@iew.unizh.ch (U. Fischbacher).

1090-5138/04/$ – see front matter
doi:10.1016/S1090-5138(04)00005-4

economics, invoke no other concept more frequently than that of "norms" (Sills, 1968, p. 208). For instance, many ethnographic descriptions contain vivid descriptions of powerful social norms, ranging from norms on mating practices and participation in religious activities to food-sharing norms and people's obligations in cooperative production or defense activities (e.g., Fessler 2002a, 2002b; Gurven, in press; Sober & Wilson 1997). Even in economics, an increasing number of researchers use this concept to explain important phenomena (e.g., Kandori, 1992; Lindbeck, Nyberg, & Weibull, 1999; Solow, 1990). Thus, it is impossible to understand human societies without an adequate understanding of social norms. Nevertheless, social norms are still poorly understood. Despite some recent progress (Hechter & Opp, 2001), we still know very little about how they are formed, the forces determining their content, how and why they change, their cognitive and emotional underpinnings, how they relate to values, how they shape our perceptions of justice and its violations, and how they are shaped by and shape our neuropsychological architecture. In short, social norms are one of the big unsolved problems in the behavioral sciences.

Social norms are also important for the evolution of human altruism because they have a bearing on the debate between individual and group selection approaches. A key argument against the empirical plausibility of group selection in the evolution of human altruism is that migration between groups removes the differences between groups. If a few selfish individuals join groups that are predominantly composed of altruistic individuals, the selfish individuals will reproduce at a higher rate, quickly removing the differences in the composition of selfish and altruistic individuals across groups. Thus, group selection cannot become operative. However, selfish migrants may not be able to reproduce at a higher rate in the presence of social norms proscribing individually selfish behavior because they are punished for violation of the norm. This means that differences in group composition can be maintained, rendering group selection empirically more plausible (Bowles, Choi, & Hopfensitz, 2003; Boyd, Gintis, Bowles, & Richerson, 2003; Gintis, 2000; Henrich & Boyd, 2001).

In this paper, we contribute to the understanding of social norms by studying underlying enforcement mechanisms. Norms are enforced due to the expectation that violations of the behavioral standard will be punished. The sanctioning individuals may be "second parties" whose economic payoff is directly affected by the norm violation. For instance, one party in an exchange relationship may violate an implicit agreement, hurting the exchange partner. The cheated partner is the "second party" in this case, while an uninvolved outside party who happens to know that cheating occurred is the "third party." Thus, the norm violation does not directly affect the third party's economic payoff. However, if only second parties imposed sanctions, a very limited number of social norms could be enforced because norm violations often do not directly hurt other people. In the case of voting norms (Knack, 1992), for example, nobody is directly hurt if somebody does not vote or votes for the "wrong" party. Likewise, in cases of cooperative effort norms, a shirking individual imposes little cost on any particular other individual if work teams are sufficiently large. Thus, third-party sanctions greatly enhance the scope for norms that regulate human behavior. In fact, some researchers view the existence of third-party sanctions as the essence of social norms because second-party punishment strategies are not evolutionarily stable in iterated pairwise interactions,

whereas strategies involving third-party sanctions are stable (Bendor & Swistak, 2001). The problem is, however, that there is still relatively little empirical knowledge about third-party sanctions at present. Much of the evidence on third-party sanctions comes from ethnographic descriptions of norm enforcement in small-scale societies (e.g., Cronk, Chagnon, & Irons, 2000; Fessler, 2002a; Sober & Wilson, 1997). Economic historians have also provided accounts of third-party punishment (Greif, 1993, 1994).

Although field evidence on the existence and enforcement of social norms is indispensable and, in fact, motivated our study, isolation of the different forces shaping norm enforcement in the field is extremely difficult if not impossible because too many uncontrolled factors simultaneously affect the results. For example, it is generally impossible in the field to distinguish between reputation-driven third-party sanctions that are motivated by selfish economic benefits and those driven by altruistic goals. However, it is possible to control for these factors in the laboratory, and we therefore examine the characteristics and relative strength of third-party sanctions in the context of laboratory experiments in this paper. We introduce, in particular, a third party into the dictator game (DG) and the prisoners' dilemma (PD) game. The third party observes the actions of the players in the DG and the PD and can then punish them. Punishment is, however, costly for the third party so a selfish third party will never punish. This design is motivated by the idea that social norms apply in both games: a norm concerning distributional fairness in the DG and a cooperation norm in the PD. The notion of strong reciprocity (Fehr & Fischbacher, 2003; Fehr, Fischbacher, & Gächter, 2002; Gintis, Bowles, Boyd, & Fehr, 2003) implies that the third parties should be willing to punish the violation of these norms. Thus, if we observe punishment of norm violations, we have evidence that strong reciprocity is also relevant for behavior of third parties.

Apart from providing direct insights into the nature of third-party sanctions, our experiments also enable us to evaluate recent theories of social preferences (Bolton & Ockenfels, 2000; Dufwenberg & Kirchsteiger, in press; Falk & Fischbacher, 1999; Fehr & Schmidt, 1999; Levine, 1998; Rabin, 1993), which assume various nonselfish preferences with diverse implications for the existence and pattern of third-party sanctions. Finally, one of the most important long-term contributions of our paper may be the provision of a simple yet powerful method for studying the characteristics and content of social norms. Whereas rewards or sanctions by second parties can often be rationalized, or are at least likely to be colored, by egocentric, ''nonnormative'' motives, the rewards and sanctions of third parties reveal the truly normative standards of behavior. For instance, norm adherence may trigger feelings of gratitude if second parties benefit from the norm adherence in economic terms and this may, in turn, induce second parties to reward those who obey the norm. In this case, the rewarding by the second party cannot be taken as unambiguous evidence for an appreciated behavioral standard. Likewise, if norm violation directly harms someone, his impulse is to retaliate, but retaliation may have nothing to do with his appreciation of behavioral standards. The study of third-party rewards and sanctions can clarify these confounding factors.

The rest of the paper is organized as follows. In the next section, we report the results of third-party sanction in the context of violations of a distribution norm. In Section 3, we deal with the case of cooperation norms, and in Sections 4 and 5, we compare the strength and the pattern of second- and third-party punishment. Section 6 briefly presents questionnaire

evidence concerning negative emotions and fairness judgments in relation to the behavioral results, and Section 7 concludes the paper.

2. Third-party sanctions of violations of a distribution norm

Many experimental economists have studied second-party sanctions in the past decade. The most famous example of such sanctions is probably the ultimatum bargaining game (Camerer, 2003; Güth, Schmittberger, & Schwarze, 1982). More recently, the study of second-party sanctions has been extended to gift exchange games (Fehr, Gächter, & Kirchsteiger, 1997), public goods games (Fehr & Gächter, 2002; Ostrom, Walker, & Gardner, 1992; Yamagishi, 1986), and taxation games (Bosman & van Winden, 2002). However, only a few papers report results on third-party sanctions in response to violations of fairness norms (Kahneman, Knetsch, & Thaler, 1986; Turillo, Folger, Lavelle, Umphress, & Gee, 2002), and we know of no paper that examines either the relative strengths of second- and third-party punishment or the pattern and strength of third-party punishment in PDs. Carpenter and Matthews (2002) studied third-party sanctions in a public goods context, but since their design allowed for reciprocity and strategic interactions among the third parties, they could not rule out third-party punishment for reasons of self-interest. As we will see below, our design completely rules out the possibility of self-interested third-party sanctions.

2.1. Methods and experiment design

We studied third-party sanctions of violations of a distribution norm by adding a third player with a punishment option to a DG played between Player A, the dictator, and Player B, the recipient. We denote this experiment as third party punishment in the dictator game (TP-DG). Player A had an endowment of 100 points and could transfer 0, 10, 20, 30, 40, or 50 to Player B, who had no endowment.[1] The third party, Player C, was endowed with 50 points, and had the option of punishing Player A after observing A's transfer to B. Player A's payoff was reduced by 3 points for every punishment point that Player C assigned to Player A. In principle, Player C could use up to 50 points (C's whole endowment) to punish A. At the end of the experiment, points were converted into real money at an exchange rate of 1 point = CHF 0.3. Player B could not affect the payoff of any other person in the game—he or she was just the passive recipient of A's transfer. However, while Player A was making his or her decision, Player B indicated the amount of punishment B expected Player C to impose on A at any feasible transfer level. In addition, B indicated how much B expected Player A to transfer to him. B's expectations were recorded by the experimenter but were never revealed to Players A and C.

[1] It is well known from many DGs (see, e.g., Camerer, 2003) that Player A almost never gives more than 50% of the available money to Player B. Therefore, to simplify the game, we did not allow Player A to transfer more than 50 points to B. In the experiment reported in Section 4, Player A could transfer more than 50 points.

Certain features were common to all the experiments reported in this paper. First, all subjects were informed about the extensive form of the game and were told each player's endowment and the exchange rate between Swiss francs and points, at the beginning of the experiment, that is, before they made their decisions or reported their expectations. Thus, Player A knew, for example, that C could punish him. Second, subjects received a show-up fee of CHF 10 (≈US$8) in all experiments; this show-up fee is not considered part of a subject's endowment, but is included when we report subjects' average earnings in the results below. Third, we never used terms like "sanction" or "punish" in the instructions, instead instructing the third parties that they had the option of assigning "deduction points" to the other players. The experimental instructions can be found in the appendix of Fehr and Fischbacher (2004). Fourth, subjects interacted anonymously and were never informed of other players' identities. Fifth, the subjects were students from the University of Zurich and the Federal Institute of Technology in Zurich. Sixth, each subject participated in only one experiment. Seventh, all experiments were based on the computer software z-Tree (Fischbacher, 1999). Eighth, a player could incur a loss in case of very severe sanctions, and it was made clear in the instructions that players had to pay their losses; in fact, however, no losses occurred.

Ninth, we implemented the so-called strategy method at the punishment stage: C had to indicate how much he or she would punish for each possible strategy combination of Players A and B. In the DG, this meant that C indicated the number of deduction points for each of A's possible transfer levels before knowing A's actual choice. The advantage of this method is that it allows analysis of sanctioning behavior in much more statistical depth; for instance, dictators rarely or never choose certain transfer levels, so if C could respond only to A's actual choice, we would have few data for those levels. A potential disadvantage of this method, however, is that it may reduce the impact of emotions: C may, for example, experience stronger emotions when reacting to an actual violation of a fairness norm than when contemplating what he would do in case of such a violation. It is, however, an open question whether the strategy method actually leads to different response probabilities than when subjects respond only to others' actual choices.[2]

Sixty-six subjects participated in the TP-DG, and each played the TP-DG only once. The roles of A, B, and C were randomly assigned to the subjects at the beginning of the experiment. The experiment lasted roughly 40 minutes and subjects earned on average CHF 22.20 (≈US$17).

The purpose of TP-DG was to see whether Player C would sanction A for violating a distribution norm. If C cares only about his or her own payoff, C should never punish, and if Player B believes C is selfish, B expects no punishment regardless of how much A

[2] It could also be argued that the strategy method dilutes the monetary incentives because subjects make more decisions for the same amount of money. However, a recent meta-study of Camerer and Hogarth (1999) indicates that the modal effect of stake size on mean experimental outcomes is zero (though variance is usually reduced by higher payment). This coincides with the results of a similar study by Smith and Walker (1993). Moreover, Brandts and Charness (2000) as well as Cason and Mui (1998) report evidence indicating that the strategy method does not induce different behaviors.

transfers. However, we hypothesized that the salient distribution norm in the DG is for A to give 50 points to B; since subjects played the game anonymously and were randomly allocated to their roles and, hence, their endowments, there is no good reason why A should end up with more money than B, making the equality norm salient. There are now several theories of social preferences (Falk & Fischbacher, 1999; Fehr & Schmidt, 1999) that are based on the behavioral relevance of equality norms, and if such norms are indeed relevant for Player C, we would expect the punishment C imposes to increase in severity the more A's transfer falls short of 50. In addition, by asking B what punishment B expects C to impose on A, we receive information on the extent to which directly affected parties expect third-party norm enforcement, and how accurate their expectations are. This information is important because the impact of social norms on behavior should increase the more people believe in the presence of third-party norm enforcement.

2.2. Results

The actual behavior of third parties disconfirms the hypothesis that they care only about their own economic payoffs and thus will never punish (Fig. 1). Most third parties punished dictators who transferred less than half their endowment, and the majority of recipients expected them to do so. At each transfer level below 50, roughly 60% ($n = 22$) of players C chose to punish the dictator A, and with the exception of transfer level 40, the proportion of recipients B who expected C to punish was higher than the proportion who actually did so.

Fig. 2 indicates that punishment and expectations thereof increased in proportion to the amount by which dictators' transfers fell short of 50%. The average punishment imposed when A gave nothing was 14 deduction points, that is, reducing A's income by 42 points, and sanctioning declined monotonically to near zero as transfers reached half of the endowment. OLS regression of punishment on the variable (50-transfer) confirms this result, yielding a highly significant ($P < .001$) slope coefficient of .28 while the constant is close to zero (-0.45) and not significant ($P = .230$).[3] Hence, although punishment for transfers of 50 did occur, its average level was not significantly different from zero. For each 10-unit reduction in points transferred, C assigned on average 2.8 deduction points, reducing A's income by 8.4; this implies that dictators who gave less than 50 gained in economic terms, but these gains were quite small.

Figs. 1 and 2 suggest that the recipients' expectations of punishment were even higher than actual punishment. If we regress B's expectations of punishment on (50-transfer), we get an insignificant constant but a highly significant slope coefficient of .33. Thus, for each reduction of the transfer by 10 units, B expected C to reduce A's income by 9.9 units. We also conducted Mann–Whitney tests to check whether these expectations

[3] Our significance tests are based on robust standard errors that take into account the fact that a given individual's punishment choices are dependent observations, whereas across individuals the punishment choices represent independent observations.

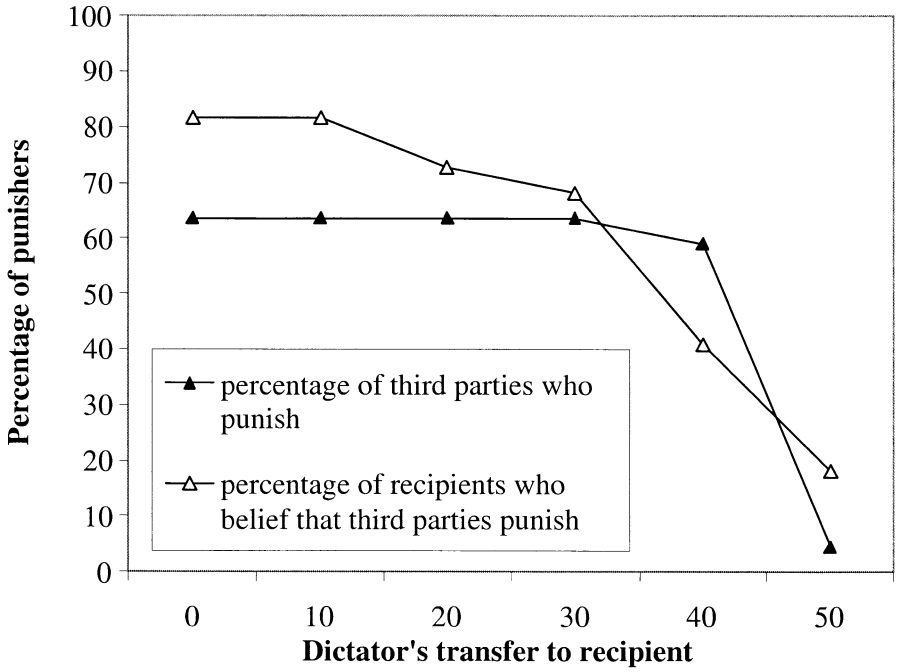

Fig. 1. Percentage of third parties who punished in the dictator game.

differed significantly from the actual punishment; these tests indicated that no significant differences exist ($P > .40$ at each transfer level below 50).

Although our primary interest is not focused on the dictators' behavior, a description of the distribution of transfers and the recipients' beliefs about them is worthwhile. The average transfer by A was 24.5 and the average expectation by B was a transfer of 21.8. A Mann–Whitney test indicates no significant difference between these values ($P = .634$). This transfer level is also quite similar to that observed in typical DGs without punishment: According to a meta-analysis by Camerer (2003), the average transfer levels across many different DG studies gravitate around 20% of the available pie. Fig. 3 also shows that the distribution of actual transfers is relatively similar to the distribution of the recipients' expectations.

2.3. Implications for proximate theories of social preferences

To what extent can recent theories of social preferences account for punishment in the TP-DG? Such theories typically assume that people are not just motivated by their own payoffs, but also care about payoffs to (relevant) others. Player C, for instance, may also care about the payoff to the dictator or the recipient in the DG. Three types of social preference theories have trouble explaining the existence of nonselfish third-party punishment. First, there are theories of altruism (e.g., Andreoni, 1989) that assume that nonselfish players care positively about the economic payoff of relevant reference actors; without incorporating additional motives (e.g., for equity), these theories never predict any punishment. Second, the theory of Bolton

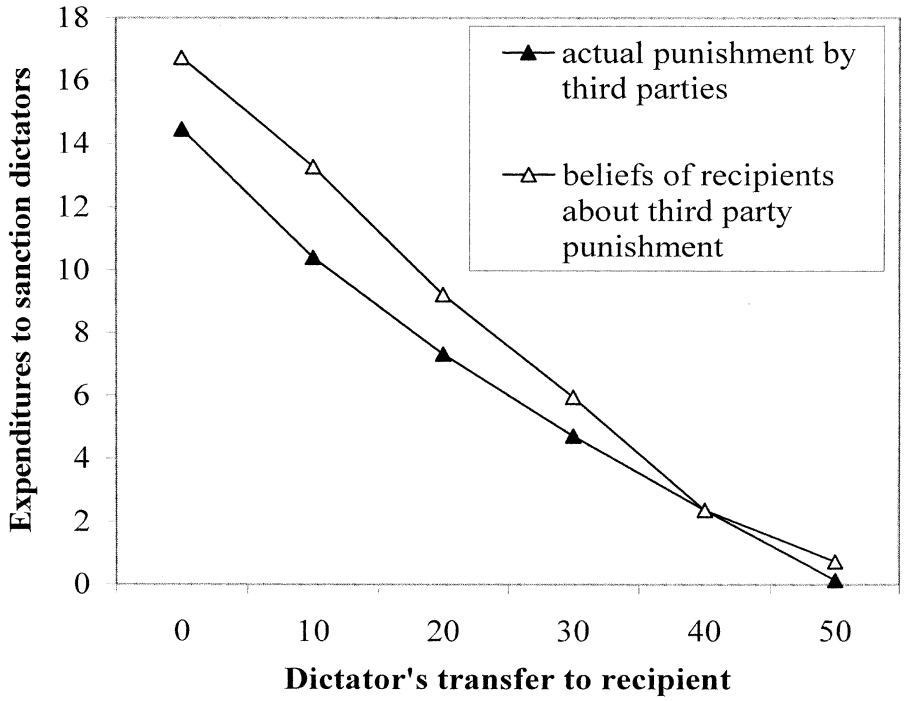

Fig. 2. Pattern of third-party punishment in the dictator game.

and Ockenfels (2000) and the pure reciprocity approach, as modeled by Rabin (1993) and Dufwenberg and Kirchsteiger (in press), have problems in explaining the evidence. Bolton and Ockenfels (BO) assume that a player is motivated by both the player's own material payoff *and* relative share of the total payoff, so if the relative share of player i is below or above the fair share defined by $1/n$ (where n is the total number of players), player i experiences a nonpecuniary disutility. Player i is therefore willing to punish in order to move his or her own share closer to the fair share. In the TP-DG, however, C was endowed with 50 points, A with 100 points, and B with 0, so regardless of what A transferred to B, C always had a "fair share" of 1/3, and BO predicts zero punishment. This failed prediction is not just an artifact of the fact that we endowed C with exactly one third. Imagine, instead, that C's endowment were only 30 points so that a C with BO preferences has less (i.e., 30/130) than a fair share; such a C should still never punish, since by spending a point on punishment C reduces his or her own payoff by 1 and the total payoff by 4, further *decreasing* his or her relative share of the total payoff to 29/126.[4] Finally, the BO approach is mute with regard to the punishment target in the TP-DG because—as long as A has transferred some money—C's

[4] To illustrate this point assume that the endowment of the third party is $x < 33.3$ and punishment is denoted by p. Then the third party's relative share after punishment is $(x - p)/(100 + x - 4p)$. Differentiating this term with respect to p yields $(3x - 100)/(100x - 4p)^2$. This derivative is negative for $x < 100/3$ so that punishment decreases the relative share of the third party.

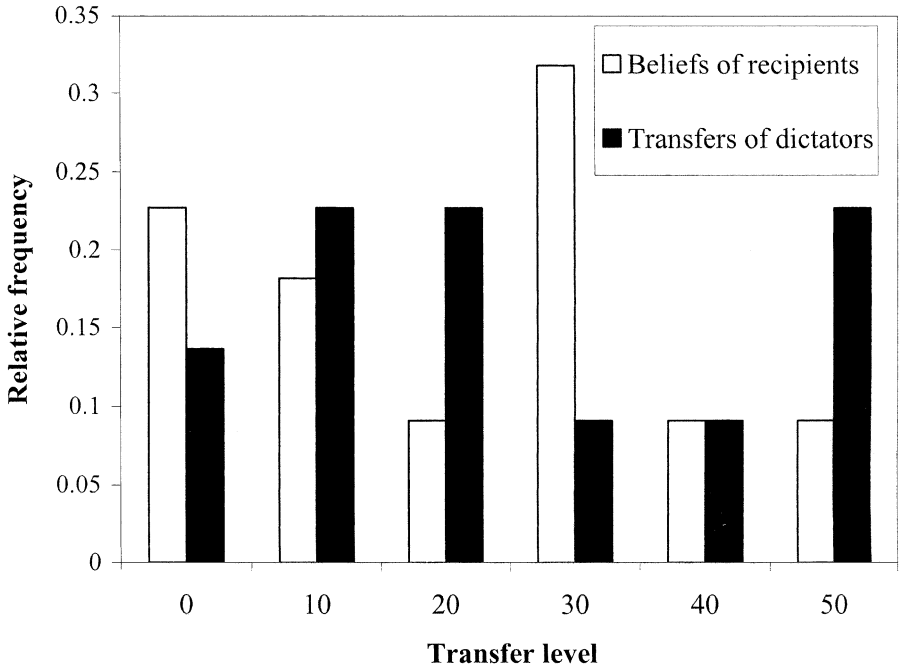

Fig. 3. Distribution of actual and expected transfers in the dictator game.

relative share can be varied regardless of whom he punishes. Thus, if we give C the option to punish A or B, the BO theory predicts that C will be indifferent between punishing A or B. The BO model therefore makes rather absurd and counterfactual predictions for the TP-DG.

The pure reciprocity approach (Dufwenberg & Kirchsteiger, in press; Rabin, 1993) rests on the assumption that unfair behavior of A towards B reveals A's unfair intentions that triggers B's willingness to punish A. However, A never behaved unfairly towards the third party in our TP games, so the pure reciprocity approach also predicts that C will never punish. In contrast, the approaches of Falk and Fischbacher (1999), Fehr and Schmidt (1999), and Levine (1998) are all consistent with the existence of third-party punishment. Levine's model rests on the assumption that there is a willingness to pay to punish individuals with selfish or spiteful preferences. In the TP-DG low transfers reveal that the dictator has such preferences, so third parties with Levine-type preferences are willing to punish dictators who transfer little.

The models of Falk and Fischbacher (1999) and Fehr and Schmidt (1999) also predict the existence of third-party punishment. The Fehr and Smith model is based on the assumption that there are players who are willing to pay to reduce the differences between their own payoffs and those of other. In the F&F model, players are willing to pay because they view inequalities as unfair; inequality is the trigger of punishment but, in contrast to the Fehr and Smith model, it is not the aim of the players to establish equality. The material payoff to C in the TP-DG is below that to A as long as the latter transfers less than the egalitarian level of 50. Thus, players with Fehr and Smith or Falk and Fischbacher preferences are willing to punish dictators who transfer less than 50 because this transfer causes a payoff difference

between themselves and the dictator. In addition, since sanctioning is costly for the third party, it also reduces the payoff difference between the third party and the recipient.

3. Third-party sanctions of violations of a cooperation norm

3.1. Methods and experiment design

Our next study assesses whether subjects will punish deviations from cooperation norms. For this purpose, we introduced a third-party punishment option into a PD game, hereafter referred to as TP-PD. As in TP-DG, the TP-PD had two decision stages. In the first, Players A and B were each endowed with 10 points and interacted with each other in a PD: each could keep his points or transfer all 10 to the other, in which case the experimenter tripled them. For example, if A transferred the 10 points while B retained his points, then B earned a total of 40 points (30 points from the transfer plus the original 10) and A earned nothing. Thus, irrespective of what the other player did, a player in the first-stage game was always better off if the player kept the endowment for himself, but if both players kept their endowments they earned only 10 points each, whereas if they both transferred their endowments each earned 30 points.

Player C observed A's and B's actions in Stage 1, and then had the opportunity to assign deduction points to A and/or B in Stage 2. Player C received an endowment of 40 points at the beginning of this stage (after not receiving any endowment in the first stage), which C could use to finance the assignment of deduction points. C could assign up to 20 deduction points to each of the two other players. As in all other experiments, assigning 1 deduction point cost C 1 point and cost the sanctioned player 3 points. To prevent the final payoff to A and B from becoming too unequal from that of C in case both A and B cooperate, we gave players A and B an additional endowment of 15 points each at the beginning of Stage 2. The additional endowment for A and B was 15 instead of 10 points because if we had given 10 points all players would have earned 40 points if A and B cooperated and C did not punish. We thought that this might create a strong focal point, and to prevent this, we gave A and B slightly more than C in this situation. As always, all players had complete knowledge of the rules of the game, every player's endowment, when the endowments were given, and C's punishment options.

Seventy-two subjects participated in the TP-PD, which lasted roughly 45 minutes. One point was worth CHF 0.37, and subjects earned on average CHF 23.40 ($\approx$ US$18.7). As in the TP-DG, the third party's decision was elicited with the strategy method. This meant that the third party indicated the number of deduction points for the sanctioned player for each of the four possible action combinations that can occur in the PD: (cc), (cd), (dc), (dd) where c stands for cooperation and d for defection. Technically, we did this by presenting C with four different computer screens, one for each of the above combinations, whereupon C indicated for each situation how many deduction points, if any, C wanted to assign to A and B.

There is considerable evidence on conditionally cooperative behaviors in public good games and prisoners' dilemmas (Dawes, 1980; Fischbacher, Gächter, & Fehr, 2001; Messick

& Brewer, 1983): Subjects are willing to cooperate if the probability that others will also do so is sufficiently large. This led us to believe that defection constitutes a much more severe norm violation if the partner in the PD cooperates than if he or she, too, defects. Therefore, we predicted that a norm of conditional cooperation shapes the pattern of third-party punishment in the PD.

3.2. Results

Table 1 shows Player C's average expenditure for punishment and the fraction of punishing C players in each possible situation. Almost half the C players (45.8%) punished the defector if the other player cooperated in the PD, and a sizeable fraction (20.8%) punished if both players defected, although the punishment was much lower than when there was only one defector. The average punishment of 3.35 deduction points imposed on a defector paired with a cooperator reflects the fact that 11 of 24 C players actually *did* punish and averaged 7.31 deduction points. Thus, third-party punishment reduced the income of a defector paired with a cooperator by $3.35 \times 3 = 10.05$ points. If, in contrast, both A and B defected, each defector received on average only 0.583 deduction points, thus losing $0.583 \times 3 = 1.75$ points. A Wilcoxon signed rank test for matched pairs shows that this difference is significant ($P = .008$), indicating that third parties perceive the same action—in our case defection—very differently, depending on what the other player in the PD did. This punishment pattern suggests that defection constitutes a less severe norm violation if the other player is also a defector.

Several features of the punishment pattern in Table 1 support the view that a considerable percentage of the players subscribe to a cooperation norm. This is indicated by the fact that mutual cooperation is almost never punished whereas mutual defection is punished in 20.8% of the cases. Moreover, even when the other player defected, defection was still more than twice as likely to be punished (20.8%) as cooperation (8.3%).

To test whether punishment in cases of mutual cooperation was significant, we regressed the amount of punishment on a dummy variable that takes on the value of 1 if the *punished* player is a defector, another dummy variable that takes on the value of 1 if the *other* player in the PD group is a defector, an interaction between the two dummies, and a constant. The results of this regression are shown in Table 2. The constant, which measures the amount of punishment if both dummies are zero, that is, if both players cooperate, is insignificant, suggesting that there is no meaningful tendency to sanction in this case. The dummy for

Table 1
Third-party punishment in the prisoners' dilemma (average expenditure)

Punished player is a	Other player in the PD-group is a defector	Other player in the PD-group is a cooperator
Defector	0.583 (20.8%)	3.354 (45.8%)
Cooperator	0.063 (8.3%)	0.083 (4.2%)

The first number in each cell denotes the average punishment of Player C. The number in parentheses denotes the percentage of Cs who punish. $n = 24$.

Table 2
Third-party punishment in the prisoners' dilemma (regression results)

	Coefficient	Robust standard error	*P* value
Punished player is a defector (Pun-def)	3.271	1.102	.007
Other player in the PD group is a defector (Other-def)	−.021	.098	.834
(Pun-def)×(Other-def)	−2.75	1.058	.016
Constant	.083	.084	.331

Dependent variable is the expenditure for sanctions by the third parties. OLS regression with clustering on individuals ($N = 192$, Prob > $F = 0.022$, adjusted $R^2 = .195$). We show robust standard errors that take into account that the sanctioning choices of a given individual in the different situations may be dependent while the sanctioning choices of different individuals are independent.

"punished player is a defector" is, however, 3.27 and highly significant. In contrast, the dummy for "other player in the PD group is a defector" is close to zero and insignificant, indicating that punishment of cooperators remains insignificant if the other player in the group changes from cooperation to defection. In other words, the punishment of a cooperator is negligible, irrespective of whether the other player cooperates or defects. This contrasts sharply with the punishment pattern for defectors. The sanctioning of a defector becomes much more severe if the other PD player changes from defection to cooperation. Finally, the negative and significant coefficient for the interaction between the dummies shows that if the punished player switches from cooperation to defection, the increase in punishment that occurs is significantly smaller if the other player in the group is a defector than if the other player is a cooperator.

These results also have implications for theories of social preferences. The existence of third-party punishment challenges the pure reciprocity approaches of Dufwenberg and Kirchsteiger (in press) and Rabin (1993), which predict no punishment in the TP-PD. The reason is again that defection in the PD implies no hostility towards Player C. Likewise, the model of Bolton and Ockenfels (2000) predicts no punishment. This can be illustrated for the case where Player A cooperates and B defects: the payoffs for A, B, and C before C's decision to punish are given by (15, 55, 40), so if C assigns one deduction point to B, C's payoff share *increases* from 40/110 = 0.364 to 39/106 = 0.368, thus moving further away from the fair share of 1/3. Therefore, a third party with Bolton and Ockenfels' preferences will not punish. This contrasts with the models of Falk and Fischbacher (1999) and Fehr and Schmidt (1999) because the payoff differences between the third party and the other players matter in these models. Thus, the third party may well punish the defector because the defector earns more than the third party. However, both models have difficulties in explaining the fact that mutual defection is also punished, because the payoff vector in this case is (25, 25, 40) before the punishment decision of C. Since Player C is better off than both A and B, Player C should never punish in this situation, but in fact, 20.8% of the third parties punished, albeit at a rather low level. The model of Levine (1998) is again consistent with the existence of third-party punishment of defectors because defection may be taken as a signal that the defector is a greedy subject.

4. Second- versus third-party punishment in the context of a distribution norm

4.1. Methods and experiment design

In the DG, second-party punishment means that the recipient, Player B, has the option of punishing the dictator. We developed the following design to compare the relative strength of second- and third-party punishment. At the beginning of the experiment, subjects were randomly assigned either the role of the dictator (Player A) or that of the recipient (Player B). Then we formed groups of two players with each group comprising one Player A and one Player B. The players in these groups then participated in a second-party punishment (SP) condition and in a third-party punishment (TP) condition according to the design described below. The sequence of the two conditions was balanced to control for order effects.

As in TP-DG, Player A was endowed with 100 points and B had no endowment in the first stage. However, Player B received an endowment of 50 points at the beginning of Stage 2 in both the SP and TP condition. To keep the payoff differences generated by A's transfer constant, we also gave A an endowment of 50 points at Stage 2. With the help of the endowment, B could finance B's sanctions even if A transferred nothing to B. In Stage 2, Player B had the option of punishing a dictator specific to the condition B is playing after observing Player A's transfer. However, while Player B could punish his dictator (Player A) in the SP condition, Player B could only punish the dictator of a different group (Player A′) in the TP condition (see Fig. 4). By comparing B's sanctioning of his own dictator (Player A) in the SP-condition to the sanctioning of the out-group dictator in the TP-condition (Player A′), we can examine the relative strength of third-party punishment. Since A′ is in a different group, A′ could not affect Player B's economic payoff. This means that with respect to the group comprising players A′ and B′, Player B was an unaffected third party. A further important feature of this treatment is that we ruled out reciprocity between the punishers; that

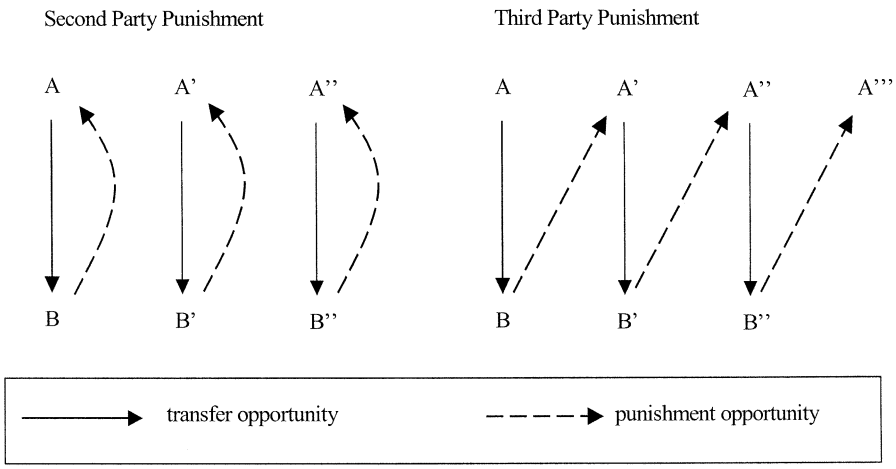

Fig. 4. Who can punish whom in the dictator game under second- and third-party punishment.

is, if Player B could punish A′ then Player B′ (who was in the group with A′) could never punish Player A (who was in the group with B). Instead, Player B′ could only punish a dictator from a third group, say A″. This feature rules out behavior that follows the principle "I punish your dictator and you punish mine." We deliberately designed the experiment in this way to ensure that the punishing players were not engaged in strategic interaction with other punishers.

In the SP condition, the punishing player is the recipient of the player's dictator's transfer and can punish him accordingly. Thus, the second party punisher experiences what it means to be the recipient of a dictator's transfer. This raises the design question of whether the punishing player in the TP condition should also be the recipient of a dictator's transfer—as is the case in our TP condition. Note that being the recipient of *another* dictator's transfer does not change the punishing player's position as a third party vis-à-vis the punished dictator. The punishing player's payoff is still unaffected by the actions of the dictator the player can punish. The choice of whether the punishing third party is placed in the role of transfer recipient depends on the question being examined. If the experience of being a transfer recipient remains constant across conditions, we can focus exclusively on the fact that the punisher is directly affected by the action of the dictator who can be punished in the second-party condition, whereas the punisher is not affected by the action of the dictator whom he can punish in the third-party condition. Since we wanted to focus on this effect, we also put the third-party punisher in the position of a transfer recipient.

When the subjects played the first condition (either SP or TP), they did not know that a second condition would follow; we first told the subjects at the end of the first condition that a second experiment would follow, and that this experiment would then be the final one. In this way, we eliminated any effect from the second condition on the first. We implemented this feature because we thought that there might be important behavioral spillovers between the first and the second conditions. In this case, by comparing the SP condition, when played first, with the TP condition, when played first, we can compare the two conditions without any confound.

In the TP condition, the third party (Player B) was informed how much B's own dictator (Player A) had transferred to him before deciding about punishing the outgroup dictator A′. The punishment decision was again elicited by the strategy method; that is, B indicated how much he would punish A′ for every possible transfer of A to B′. Likewise, in the SP condition B indicated how much he would punish A for every possible transfer. There is thus a difference between the SP and TP conditions: The punishing player in the TP condition knew how much he had received from his own dictator before his punishment decision. We introduced this feature because if we did not tell third parties in the TP condition how much they had received, they would have had beliefs about this transfer anyway. Thus, only two possibilities were available: controlling the beliefs of the third parties by telling them how much they had received or accepting an uncontrolled belief. We decided that it would be better to know the third parties' beliefs and to use this information as a control variable in our statistical analysis (see below).

In this experiment, which lasted roughly 75 minutes, 92 subjects participated in both the SP and TP conditions. The exchange rate was 1 point = 0.13 CHF, and average earnings were CHF 30.

4.2. Results

Fig. 5 presents the main results of this experiment. Dictators faced severe sanctions in both the second- and third-party conditions, but second-party sanctions for transfers below the egalitarian level were considerably stronger than those by third parties, with the effect that low transfers were profitable for dictators in the TP condition but not in the SP condition. The figure shows that second parties punished more than third parties for all transfer levels below 50, while punishment was generally very low and similar across conditions for transfer levels above 50. These qualitative differences between second- and third-party punishment were the same regardless of whether the SP or TP condition was conducted first.

Dictators were strongly sanctioned in both conditions. In fact, the punishment for transfers below 50 in the SP condition was so high that dictators always earned less money if they gave less than 50. Fig. 6 portrays the dictators' expected payoff (i.e., average earnings minus average punishment costs) for each transfer level and shows that the egalitarian transfer maximized the dictators' payoff in the SP condition. In the TP condition, the situation was different: A transfer of 10 was more profitable than the other transfer levels.[5]

To test whether differences between second- and third-party punishment were significant, we ran OLS regressions with robust standard errors (see Table 3). As in our previous regression-based tests, these standard errors take into account that only the observations across individuals are independent, whereas different choices of given individuals are not. In the regressions, we only used the data from those SP and TP conditions that were conducted as first conditions in a session because we detected spillover effects across SP and TP conditions. The punishment level was significantly higher (3.5 deduction points) when the TP condition was conducted as the first condition than when it was conducted second, indicating that there was a spillover effect from the SP to the TP condition. To keep this spillover from contaminating our statistical results, the regressions in Table 3 are based only on data from the first condition in a session. Since third-party punishment was higher when the TP condition was conducted first, relying on these data makes it more difficult to detect differences between the SP and TP conditions.

The regression in the first column of Table 3 is based on data from the SP condition only. Punishment by second parties is regressed on the variable D_{neg}, which is defined as the maximum of the two numbers (0, 50-transfer), and on the variable D_{pos}, which is defined as the maximum of the two numbers (0, transfer-50). D_{neg} measures the negative deviation from the egalitarian transfer, that is, by how much a given transfer is below 50. D_{pos} measures the positive deviation of a given transfer from the egalitarian level. Notice that if D_{neg} is positive D_{pos} must be zero and vice versa. This specification is suggested by Fig. 6, which clearly shows that punishment responded differently to transfer levels below or above

[5] Some dictators seem to have anticipated this difference in punishment across treatments. While the modal transfer level is zero in the TP condition, the modal transfer level is 50 in the SP condition. A Wilcoxon signed rank test for matched pairs shows, however, that—despite the shift in the modal offer—the average offer is not different across conditions. This suggests that, to have an impact on the dictators' behavior, dictators have to experience that low transfers do not pay in the SP condition.

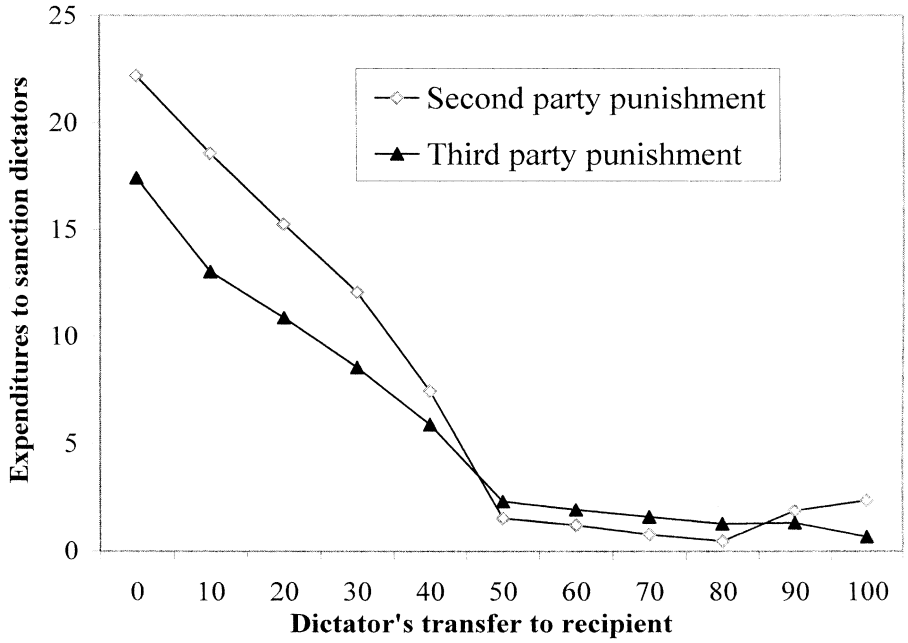

Fig. 5. Comparison of second- and third-party punishment in the dictator game.

50, respectively. A further advantage of this regression is that the constant measures punishment at the egalitarian transfer level.

The regression for the SP condition in Table 3 shows that both the constant and D_{pos} are insignificant, indicating that punishment was negligible at the egalitarian transfer level and remained so for transfers above 50. However, an increase in D_{neg} of 10 units increased punishment by 4.54 deductions points for transfers at or below 50, reducing the dictator's income by 13.62 points, confirming that deviations from egalitarian transfer were not profitable for dictators in the SP condition. In the second column of Table 3, we show the regression for the TP condition. We use again D_{neg} and D_{pos} as regressors, but we also add the transfer that third parties received from their own dictators as an explanatory variable. However, the coefficient on "transfer to third party" is low and insignificant. Likewise, the constant is insignificant in the TP regression. For transfer levels at or below 50, however, an increase in D_{neg} significantly increased punishment by third parties.

A comparison of coefficients in the two conditions shows that D_{neg} had a stronger impact in the SP condition. To assess whether this difference was significant, we ran a regression with the data from both conditions (see column three in Table 3). We added a dummy for the TP condition in this regression, and interacted this dummy with D_{neg} and D_{pos}. The regression shows that the TP dummy is insignificant, suggesting that the punishment level was not significantly different across conditions at the egalitarian transfer. The coefficient for the interaction term $D_{neg}\times$TP Dummy is significantly negative, however, indicating that punishment was less severe in the TP condition for transfer levels below 50. The small

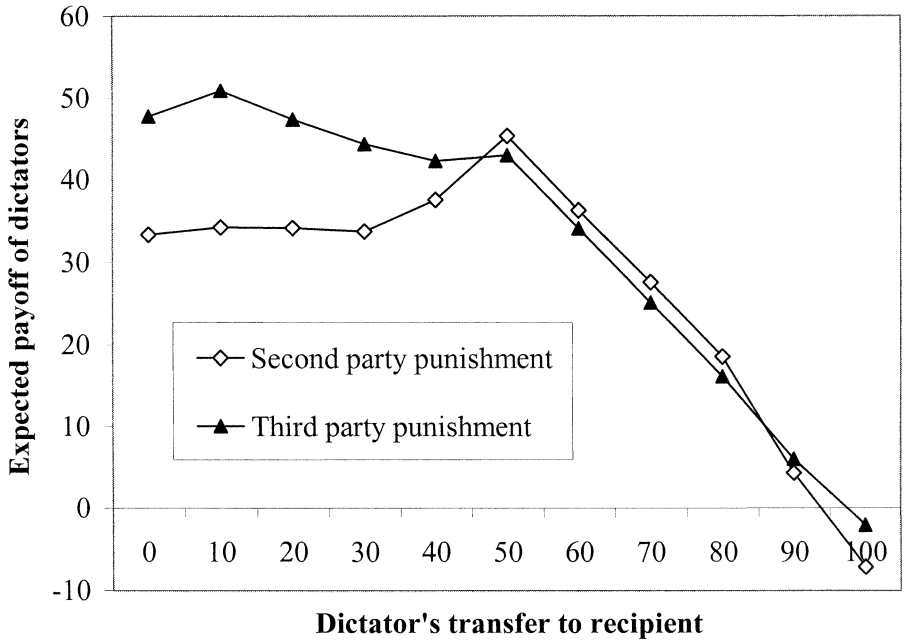

Fig. 6. Expected payoffs of dictators under second- and third-party punishment.

and insignificant coefficient for the D_{pos}×TP Dummy interaction confirms that punishment remained negligible, and did not differ across conditions, for transfers above 50.

Thus far, we have examined the aggregate pattern of punishment in both conditions, but what about individual differences? When we examine individual sanctioning patterns, it turns

Table 3
Relative strength of third-party punishment in the dictator game

	Second-party punishment	Third-party punishment	Second- and third-party punishment
Constant	1.676 (0.999)	2.076 (2.241)	1.676 (0.999)
D_{neg}	0.454*** (0.087)	0.207*** (0.061)	0.454*** (0.087)
D_{pos}	−0.025 (0.026)	−0.070*** (0.024)	−0.025 (0.026)
Transfer to third party		0.106 (0.085)	
TP-dummy			2.624 (1.667)
D_{neg} × TP-dummy			−0.247** (.105)
D_{pos} × TP-dummy			−0.046 (0.035)
No. of observations	264	242	506
Prob > F	.0001	.016	.0000
Adjusted R^2	.373	.211	.309

OLS regressions with clustering on individuals. Robust standard errors in parentheses.
** Denotes significance at the 5% level.
*** Denotes significance at the 1% level.

out that there are four clusters of people. Some subjects never punished. Another group punished only transfer levels below 50; sanctioning by those in this category generally increased monotonically in D_{neg}. Thirdly, there were subjects whose sanctions likewise increased monotonically with D_{neg} but who also punished some transfers of more than 50; in general, their sanctions for transfers above 50 were low and applied to some but not all such transfers. Finally, a few subjects exhibited rather peculiar sanctioning patterns that did not fall into any of the above categories. Two subjects, for example, imposed one deduction point on the dictator at all feasible transfer levels; three others punished those who gave little and those who gave (close to) everything but not the intermediate transfers. Table 4 shows the percentage of subjects in each of these categories across conditions. In the SP condition, 26% never punished and 39% only punished transfers below 50. This is reversed in the TP condition, indicating that the lower average sanctions for transfers below 50, relative to the SP condition, were also the result of a smaller number of punishers. The fraction of monotone punishers who also punished some transfers levels above 50 was 20% and 26%, respectively.

Comparison between the SP and TP conditions also has implications for the different social preference theories. Since we know already that the model by Bolton and Ockenfels (2000) and the pure reciprocity models fail to capture third-party punishment, we concentrate on the other models. The model by Levine predicts no difference between second- and third-party punishment, because a given low transfer to the recipient reveals the dictator's selfish or spiteful preferences regardless of whether the recipient (i.e., the second party) or the third party can punish. Thus, Levine's model cannot explain the treatment differences. This contrasts with the Fehr–Schmidt and Falk–Fischbacher models, which predict that third parties will punish less than second parties. The payoff difference between the dictator and the potential punisher increases by 2 units under the SP condition for any additional unit kept by the dictator; in the TP condition the payoff difference increases by only 1 unit. This follows simply from the fact that the potential punisher in the SP condition is also the recipient in a DG. Therefore, any additional unit the dictator keeps below the egalitarian transfer level induces more punishment in the SP condition than in the TP condition. The intuition behind this prediction is simply that the nonpecuniary harm for the potential punisher created by greedy transfers is higher in the SP condition than in the TP condition.

This prediction coincides with the predictions of a more psychological approach that stipulates that the anger experienced due to a certain action is given by the level of arousal times the salience of the cue. In our case, the transfer level by the dictator measures the

Table 4
Classification of subjects according to their sanctioning pattern

Type of subject	Second-party punishment (n = 46) (%)	Third-party punishment (n = 46) (%)
Never punish	26	39
Punish only transfers below 50	39	26
Monotone punishers who punish also some transfers at or above 50	20	26
Others	15	9

salience of the cue, whereas the level of arousal (for a given salience of the cue) is determined by how much it hurts psychologically to receive a low transfer. Since a low transfer is more harmful in the SP condition it seems reasonable to assume that the level of arousal is also larger. Thus, according to this view, subjects will experience more anger in the SP condition, which induces them to punish more severely than in the TP condition.

There is one aspect of the data that neither the Fehr–Schmidt nor Falk–Fischbacher model predicts satisfactorily. Recall that third-party punishment was not significantly affected by the sum received from one's own dictator (see regression 2 in Table 3). This contradicts both models because the more a third party receives from his or her own dictator, the less reason there is for him or her to punish the dictator in the other group, since the income difference between the third party and the dictator in the other group becomes smaller if one's own dictator transfers more. The insignificant impact of one's own dictator's transfer on third-party punishment is consistent with Levine's model, however, because the behavior of one's own dictator does not affect inferences about the dictator's preferences in the other group.

5. Second versus third-party punishment in the context of cooperation norms

5.1. Methods and experiment design

In order to compare second- and third-party punishment in the PD, we used a similar design to that in the previous section (see Fig. 7). Subjects were randomly assigned to two player groups who played a PD at the first decision stage. This PD was identical to the one in TP-PD. At the beginning of the second decision stage, each player received an additional endowment of 20 points. Then Players A and B could sanction each other in the SP condition, while Player B could only sanction a Player A′ from another group and Player A could only sanction a Player B″ from a third group in the TP condition. Again, we implemented the

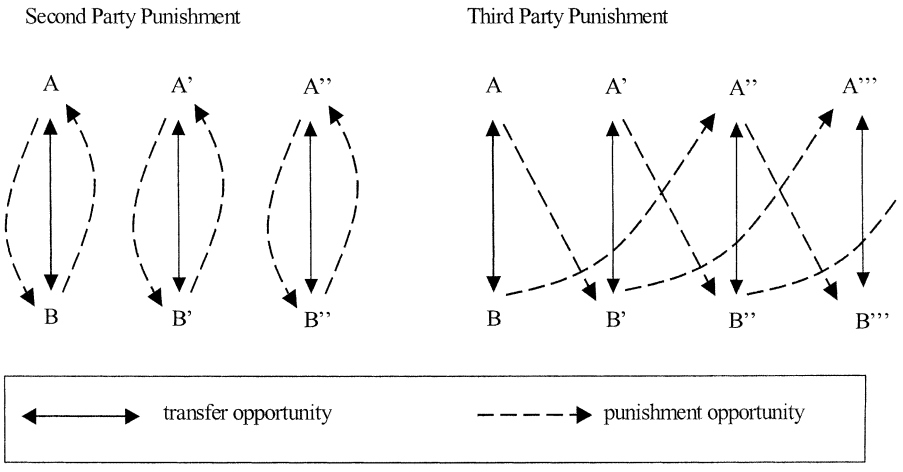

Fig. 7. Who can punish whom in the prisoners' dilemma under second- and third-party punishment.

strategy method in the punishment stage and ruled out reciprocity between the punishers in the TP condition. Note that in order to compare, ceteris paribus, second-party punishment with third-party punishment, it is inevitable that the punishing player can only sanction *one* other player—the other player in his own group in the SP condition or a member of another group in the TP condition. This differs from TP-PD, because Player C could punish both PD-players in this experiment. As usual, subjects were fully informed about the rules of the game.

Ninety-four subjects participated in this experiment, which lasted roughly 75 minutes. Each subject participated in both a TP and an SP condition and with half experiencing each sequence of conditions. The exchange rate was 1 point = CHF 0.3, and average earnings were CHF 33.40.

5.2. Results

The results presented and analyzed here are based only on data from the first condition in each experiment because, as in the case of the DG, there were strong effects of the order of conditions. The main results were as follows: both second and third parties strongly punished defectors whereas punishment of cooperators was negligible, but second-party punishment was much stronger than third-party punishment, with the result that defection was profitable in the TP but not in the SP condition (see Tables 5 and 6). Regardless of treatment condition and whether the PD partner of the punished subject defected or cooperated in the TP condition, defectors were always punished significantly more than cooperators (Wilcoxon signed rank test, $P < .0001$ in all cases). Table 6 further supports this conclusion by showing that the percentage of subjects who punished a defector varied—depending on the condition—between 32.6% and 66.7%, while the percentage of subjects who punished a cooperator varied between 8.3% and 15.2%. The greater punishment in the SP than in the TP condition was also significant, regardless of whether the punished defector's PD partner cooperated or defected ($P < .01$ by Mann–Whitney test in both cases).

The numbers in Tables 5 and 6 imply that defection was not profitable for the defector in the SP condition: the defector's average income was reduced by $3 \times 8.4 = 25.2$ points, whereas the gain from defection was only 10, generating a net loss of 15.2 points. Therefore, cooperation was the better choice from a purely monetary viewpoint. (The small residual punishment imposed on a cooperator in the SP condition caused an income reduction of only $3 \times 0.67 = 2.01$ points.) In the TP condition, however, a defector whose

Table 5
Relative strength of third-party punishment in the prisoners' dilemma (average expenditure of third and second parties for punishment)

Punished person is a	Second-party punishment	Third-party punishment if the punished person's PD partner cooperates	Third-party punishment if the punished person's PD partner defects
Defector	8.40	3.09	1.43
Cooperator	0.67	0.59	0.69

Table 6
Relative frequency of punishing individuals in the prisoners' dilemma

Punished person is a	Second party punishment	Third party punishment if the punished person's PD partner cooperates	Third party punishment if the punished person's PD partner defects
Defector	66.7% (32 out of 48)	58.7% (27 out of 46)	32.6% (15 out of 46)
Cooperator	8.3% (4 out of 48)	15.2% (7 out of 46)	15.2% (7 out of 46)

PD partner cooperated incurred a cost of $3\times3.09 = 9.27$, and defection led to an overall income of $30+10-9.27 = 30.73$, whereas a cooperative choice led to a small residual punishment cost of $3\times0.59 = 1.77$ and an overall income of $30-1.77 = 28.33$. Qualitatively, the same argument holds where a defector's PD partner also defected because third-party sanctions were even weaker in this case. Thus, given the punishment pattern, defection in the TP condition was profitable regardless of the PD partner's actions.

Our previous emphasis has been on how much potential punishers are willing to pay to punish defectors relative to cooperators. Another question is whether the willingness to incur costs for punishing defectors depends significantly on whether the punisher himself cooperated or defected. This question is particularly interesting in the TP condition because a cooperator has no direct reason to feel exploited by the defection of an *outside* group member in this condition. However, those subjects who cooperate themselves are perhaps better able to empathize with cooperators in other groups who had to face the defection of their PD partner.

Tables 7 and 8 indicate that both cooperators and defectors in the PD punished defectors, but cooperators punished more frequently and imposed much stronger sanctions[6]. Table 7 splits up the first row of Table 5, and Table 8 splits up the first row of Table 6, according to whether the punisher was a cooperator or a defector. Table 7 shows that cooperators punished defectors considerably more than did defectors, regardless of whether the sanctioning subject was a second or a third party. These results are also supported by Table 8, indicating that the percentage of punishing cooperators was considerably higher than the percentage of punishing defectors; among the defectors, however, 27% to 50% also punished other defectors. The differing willingness of cooperators and defectors to punish was marginally significant in both the SP condition ($P = .068$, Mann–Whitney test) and the TP condition ($P = .081$) if the punished defector's PD partner cooperated. The difference

[6] Intuitively, the presence of punishing defectors may be surprising, but there are several plausible reasons why defectors may punish other defectors or even cooperators. Spiteful or competitive individuals (''inequity lovers'') who aim at maximizing the difference between their own payoff and that of the other players punish regardless of whether the other person cooperated or defected. Self-serving biases may also be a reason. Assume, for instance, that a third party defects in his own group because he is a conditional cooperator who only cooperates if he believes that the PD partner will do so. Yet, he has pessimistic beliefs about the PD partner and, therefore, he defects. This third party might nevertheless punish a defector in another group if the defector's PD partner cooperated because he overlooks (or discounts the possibility) that the defector in the other group also might have been a conditional cooperator with pessimistic beliefs.

Table 7
Who sanctions defectors (average expenditure)

Punisher is a	Punished person is a	Second-party punishment	Third-party punishment if the punished person's PD partner cooperates	Third-party punishment if the punished person's PD partner defects
Cooperator	Defector	9.21	3.65	1.71
Defector	Defector	2.67	1.93	0.87

was not significant, however, if the punished defector's PD partner also defected ($P = .430$). (Note that the number of defectors was relatively small, rendering it difficult to attain high significance levels.)

Table 7 enables us to illustrate a potential objection to our comparison of second party and third-party sanctions. In Tables 5 and 6, we did not hold the action of the potential punishers in their own PD group constant across conditions. This follows necessarily from the fact that we did not specify the actions of the punishers in their own PD group in Table 5. Thus we did not test whether potential punishers *who cooperated in their own PD group* punished a defector in the SP condition more than one in the TP condition. Yet, it becomes clear that this is indeed the case upon examination of Table 7. A cooperator in the SP condition spent 9.21 points for punishing a defector, whereas a cooperator in the TP condition spent "only" 3.65 points to punish a defector who faced a cooperating PD partner. This difference is significant ($P = .007$, Mann–Whitney test) and contrasts with the case of punishing defectors. Subjects who defected in their own group spent 2.67 points to punish a defector in the SP condition and 1.93 points in the TP condition (in case the PD partner cooperated). This difference is not significant ($P > .3$, Mann–Whitney test). Thus, cooperators punished defectors much more in the SP condition than in the TP condition, whereas defectors punish defectors to roughly the same extent across conditions.

6. Negative emotions and fairness judgments

Influential social scientists (Elster, 1989; Frank, 1988; Hirshleifer, 1987) have argued that the sanctions that enforce social norms are based on strong emotions, and that emotions are the drivers of norm enforcement decisions. Moreover, Elster (1989) argued that being the object of negative emotions such as anger causes a large disutility on its own, independent of any material losses. Therefore, whether cooperating or defecting subjects anticipate the

Table 8
Who sanctions defectors (relative frequency)

Punisher is a	Punished person is a	Second-party punishment	Third-party punishment if the punished person's PD partner cooperates	Third-party punishment if the punished person's PD partner defects
Cooperator	Defector	69% (29 out of 42)	67.7% (21 out of 31)	35.5% (11 out of 31)
Defector	Defector	50% (3 out of 6)	40% (6 out of 15)	26.7% (4 out of 15)

emotions triggered by their behavior is of interest. All subjects who participated in the TP condition of the PD game filled out a questionnaire after the experiment designed to elicit self-reports about third parties' positive and negative emotions in different hypothetical scenarios.

For brevity, we do not report our questionnaire results (interested readers may consult Fehr & Fischbacher, 2004), but they are consistent with the view that emotions cause the sanctioning decisions that enforce social norms, because the pattern of third-party punishment and the emotional pattern fit together nicely. However, the results do not prove that emotions cause the sanctioning decisions. We are aware that self-reported or predicted emotions need to be treated with caution. It is possible that self-reported emotions deviate systematically from the emotions that subjects actually would experience if they were in a particular scenario. However, our questionnaire results suggest at least interesting hypotheses regarding the role of emotions in norm enforcement.

7. Concluding remarks

In this paper, we studied the enforcement mechanisms behind social norms, finding that a large percentage of subjects are willing to enforce distribution and cooperation norms even though they incur costs and reap no economic benefit from their sanctions and even though they have not been directly harmed by the norm violation. Thus, third-party sanctions provide a further important example for the notion of strong reciprocity (Fehr & Fischbacher 2003; Fehr et al., 2002; Gintis et al., 2003). Our questionnaire results are consistent with the view that third-party sanctions are driven by negative emotions and negative fairness judgments towards norm violators.

We also found that sanctions by second parties directly harmed were much stronger than third-party sanctions, indeed strong enough to make norm violations unprofitable, whereas the sanctions of a single third party were not. Thus, in the context of our experiment, more than one third party is needed to enforce the norm. However, this condition is probably met frequently in real life. Therefore, taken together, our results suggest that altruistic third-party sanctions are likely to be powerful enforcers of social norms.

We also believe that our experiments can be a useful tool for study of the content and strength of distribution and cooperation norms across different societies and cultures. It is, for instance, well known that many small-scale societies are characterized by food-sharing norms (Gurven, in press). Our third-party punishment experiments may be a useful instrument for examining the forces behind food-sharing and other cooperative activities. Finally, we believe that the experiments can form the basis for the construction of convincing proximate and ultimate models of human altruism. Such theories should be able to explain the prevailing patterns of third-party punishment and the punishment patterns associated with the evolution of social norms. Proximate models of pure strong reciprocity (Dufwenberg & Kirchsteiger, in press; Rabin, 1993) as well as the inequity aversion model by Bolton and Ockenfels (2000) are clearly unable to account for our data. In contrast, the models by Falk and Fischbacher (1999), Fehr and Schmidt (1999), and Levine (1998) predict most, but not all, aspects of our data quite well.

References

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy*, *97*, 1447–1458.

Bendor, J., & Swistak, P. (2001). The evolution of norms. *American Journal of Sociology*, *106*, 1493–1545.

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

Bosman, R., & van Winden, F. (2002). Emotional hazard in a power-to-take experiment. *Economic Journal*, *112*, 147–169.

Bowles, S., Choi, J.-K., & Hopfensitz, A. (2003). The co-evolution of individual behaviours and social institutions. *Journal of Theoretical Biology*, *223*, 135–147.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences USA*, *100*, 3531–3535.

Brandts, J., & Charness, G. (2000). Hot versus cold: Sequential responses and preference stability in experimental games. *Experimental Economics*, *3*, 227–238.

Camerer, C. (2003). *Behavioral game theory*. Princeton, NJ: Princeton University Press.

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

Carpenter, J. P., & Matthews, P. H. (2002). *Social reciprocity* (Mimeo). Middlebury, VT: Middlebury College.

Cason, T., & Mui, V. (1998). Social influence in the sequential dictator game. *Journal of Mathematical Psychology*, *42*, 248–265.

Cronk, L., Chagnon N. A., Irons W. (Eds.) (2000). *Adaptation and human behavior: An anthropological perspective*. Hawthorne, NY: de Gruyter.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, *31*, 169–193.

Dufwenberg, M. & Kirchsteiger, G. (in press). A theory of sequential reciprocity. *Games and Economic Behavior*.

Elster, J. (1989). The cement of society—A study of social order. Cambridge: Cambridge University Press.

Falk, A., & Fischbacher, U. (1999). *A theory of reciprocity*, Zurich, Switzerland: University of Zürich, Institute for Empirical Research in Economics.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*, 785–791.

Fehr, E. & Fischbacher U. (2004). *Third party punishment and social norms* (Working Paper No. 106). Zürich, Switzerland: University of Zürich, Institute for Empirical Research in Economics.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation and the enforcement of social norms. *Human Nature*, *13*, 1–25.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.

Fehr, E., Gächter, S., & Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device. *Econometrica*, *65*, 833–860.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition and cooperation. *Quarterly Journal of Economics*, *114*, 817–868.

Fessler, D. (2002a). Windfall and socially distributed willpower: The psychocultural dynamics of rotating Savings and Credit Associations in a Bengkulu Village. *Ethos*, *30*, 25–48.

Fessler, D. (2002b). *Third-party attitudes towards incest: Evidence for the Westermarck effect* (Working Paper). Los Angeles, CA: University of California Los Angeles, Department of Anthropology.

Fischbacher, U. (1999). *Z-tree—Zürich toolbox for readymade economic experiments* (Working Paper No. 21). Zürich, Switzerland: University of Zürich, Institute for Empirical Research in Economics.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative?—Evidence from a public goods experiment. *Economic Letters*, *71*, 397–404.

Frank, R. (1988). *Passions within reason—The strategic role of emotions*. New York: W. W. Norton.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*, 169–179.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution & Human Behavior*, *24*, 153–172.

Greif, A. (1993). Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *American Economic Review*, *83*, 525–548.

Greif, A. (1994). Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy*, *102*, 912–950.

Gurven, M. (in press). To give or give not: The behavioural ecology of human food transfers. *Behavioral and Brain Sciences*.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, *3*, 367–388.

Hechter, M., & Opp, K. D. (2001). *Social norms*. New York: Russell Sage Foundation.

Henrich, J., & Boyd, R. (2001). Why people punish defectors-weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*, 79–89.

Hirshleifer, J. (1987). On the Emotions as Guarantors of Threats and Promises. In John Dupre, ed., *The Latest on the Best: Essays on Evolution and Optimality,* Cambridge, MA: Bradford Books; The MIT Press.

Kandori, M. (1992). Social norms and community enforcement. *Review of Economic Studies*, *59*, 63–80.

Kahneman, D., Knetsch, J., & Thaler, R. (1986). Fairness as a constraint on profit-seeking. *American Economic Review*, *76*, 728–741.

Knack, S. (1992). Civic norms, social sanctions, and voter turnout. *Rationality and Society*, *4*, 133–156.

Levine, D. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, *1*, 593–622.

Lindbeck, A., Nyberg, S., & Weibull, J. (1999). Social norms and economic incentives in the welfare state. *Quarterly Journal of Economics*, *16*, 1–35.

Messick, D., & Brewer, M. (1983). Solving social dilemmas—A review. In L. Wheeler (Ed.), *Review of personality and social psychology.* Beverly Hills, CA: Sage.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review*, *86*, 404–417.

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

Sills, D. L. (Ed.) (1968). *International encyclopedia of the social sciences*. New York: Macmillan.

Smith, V. L., & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry*, *31*, 245–261.

Sober, E., & Wilson, D. S. (1997). Unto others—The evolution and psychology of unselfish behavior. Cambridge, MA: Harvard University Press.

Solow, R. (1990). *The labor market as a social institution*. Oxford: Basil Blackwell.

Turillo, C. J., Folger, R., Lavelle, J. J., Umphress, E. E., & Gee, J. O. (2002). Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. *Organizational Behavior and Human Decision Processes*, *89*, 839–865.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*, 110–116.