Cooperation and Punishment in Public Goods Experiments

By Ernst Fehr and Simon Gächter*

Casual evidence as well as daily experience suggest that many people have a strong aversion against being the "sucker" in social dilemma situations. As a consequence, those who cooperate may be willing to punish free-riding, even if this is costly for them and even if they cannot expect future benefits from their punishment activities. A main purpose of this paper is to show experimentally that there is indeed a widespread willingness of the cooperators to punish the free-riders. Our results indicate that this holds true even if punishment is costly and does not provide any material benefits for the punisher. In addition, we provide evidence that free-riders are punished the more heavily the more they deviate from the cooperation levels of the cooperators. Potential free-riders, therefore, can avoid or at least reduce punishment by increasing their cooperation levels. This, in turn, suggests that in the presence of punishment opportunities there will be less free riding. Testing this conjecture is the other major aim of our paper.

* Institute for Empirical Research in Economics, University of Zurich, Blümlisalpstrasse 10, CH-8006 Zurich (e-mail: efehr@iew.unizh.ch; gaechter@iew.unizh.ch; website: http://www.unizh.ch/iew/grp/fehr/index.html). This paper is part of the EU-TMR Research Network ENDEAR (FMRX-CT98-0238). Fehr also acknowledges the hospitality of the Center for Economic Studies in Munich and support from the MacArthur Foundation Network on Economic Environments and the Evolution of Individual Preferences and Social Norms. Part of the experiments are also financed by the Swiss National Science Foundation under Project No. 1214-051000.97. We gratefully acknowledge valuable comments by two anonymous referees, seminar participants at the MacArthur Foundation Meeting in Stanford, the Workshop in Experimental Economics in Berlin, the ASSA Meeting in New York, the IAREP conference in Valencia, the Econometric Society European Meeting in Toulouse, the ESA meeting in Mannheim, and the European Economic Association conference in Berlin; and by seminar participants at the universities of Basel, Bern, Bonn, Dortmund, Lausanne, Linz, Munich, Pittsburgh, St. Gallen, and Tilburg; and by Richard Beil, Samuel Bowles, Robert Boyd, Martin Brown, Robyn Dawes, Armin Falk, Urs Fischbacher, Herbert Gintis, John Kagel, Georg Kirchsteiger, Serge Kolm, David Laibson, George Loewenstein, Tanga McDaniel, John Miller, Paul Romer, and Klaus Schmidt. We are particularly grateful to Urs Fischbacher who did the programming.

For this purpose we conducted a public good experiment with and without punishment opportunities. In the treatment without punishment opportunities *complete* free-riding is a dominant strategy. In the treatment with punishment opportunities punishing is costly for the punisher. Therefore, purely selfish subjects will never punish in a one-shot context. This means that if there are only selfish subjects, as is commonly assumed in economics, the treatment with punishment opportunities should generate the same contribution behavior as the treatment without such opportunities. The reason is, of course, that the presence of punishment opportunities is irrelevant for the contribution behavior if there is no punishment. In sharp contrast to this prediction we observe vastly different contributions in the two conditions. In the no-punishment condition contributions converge to very low levels. In the punishment condition, however, average contribution rates between 50 and 95 percent of the endowment can be maintained.

The strong regularities observed in our experiments suggest that powerful motives drive the punishment of free-riders. In our view this motive is likely to play a role in many social interactions, such as industrial disputes, in team production settings, or, quite generally, in the maintenance of social norms. If, for example, striking workers ostracize strike breakers (Hywel Francis, 1985) or if, under a piece rate system, the violators of production quotas are punished by those who stick to the norm (e.g., F. J. Roethlisberger and W. J. Dickson, 1947), it seems likely that similar forces are at work as in our experiments.[1]

Our work is most akin to the seminal paper

[1] Francis's (1985 p. 269) description of social ostracism in the communities of the British miners provides a particularly vivid example. During the 1984 strike of the miners, which lasted for several months, he observed the following: "To isolate those who supported the 'scab union,' cinemas and shops were boycotted, there were expulsions from football teams, bands and choirs and 'scabs' were compelled to sing on their own in their chapel services. 'Scabs' witnessed their own 'death' in communities which no longer accepted them."

TABLE 1—TREATMENT CONDITIONS

	Stranger-treatment Random group composition in each period (Sessions 1–3)	Partner-treatment Group composition constant across periods (Sessions 4 and 5)
Without punishment (ten periods)	18 groups of size n	10 groups of size n
With punishment (ten periods)	18 groups of size n	10 groups of size n

by Elinor Ostrom et al. (1992). These authors allowed for costly punishment in a repeated common pool resource game. However, in their experiments the *same* group of subjects interacted for an *ex ante unknown* number of periods, and subjects could develop an *individual* reputation. Hence, there were material incentives for cooperation and for punishment. To rule out such material incentives we eliminated all possibilities for individual reputation formation and implemented treatment conditions with an *ex ante known* finite horizon. In addition, we also had treatments in which the group composition changed randomly from period to period, and treatments in which subjects met only once.

Our work is also related to the interesting study of David Hirshleifer and Eric Rasmusen (1989) who show that, if there are opportunities for ostracizing noncooperators, rational egoists can maintain cooperation for $T - 1$ periods in a T-period prisoner's dilemma. In this model ostracizing noncooperators is part of a subgame-perfect equilibrium and thus rational for selfish group members. This feature distinguishes the preceding model from our experimental setup. In our experiments cooperation or punishment can never be part of a subgame-perfect equilibrium if rationality and selfishness are common knowledge. We deliberately designed our experiments in this way to examine whether people punish free-riders even if it is against their material self-interest.

I. The Experimental Design

A. *Basic Design*

Our overall design consists of a public good experiment with four treatment conditions (see Table 1).[2] There is a "Stranger"-treatment with *and* without punishment opportunities and a "Partner"-treatment with *and* without punishment opportunities. In the Partner-treatment the same group of $n = 4$ subjects plays a finitely repeated public good game for ten periods, that is, the group composition does not change across periods. Ten groups of size $n = 4$ participated in the Partner-treatment. In contrast, in the Stranger-treatment the total number of participants in an experimental session, $N = 24$, is randomly partitioned into smaller groups of size $n = 4$ in each of the ten periods. Thus, the group composition in the Stranger-treatment is randomly changed from period to period.[3] The treatment without punishment opportunities serves as a control for the treatment with punishment opportunities. In a given session of the Stranger-treatment the *same N* subjects play ten periods in the punishment and ten periods in the no-punishment condition. Similarly, in a session of the Partner-treatment all groups of size n play the punishment and the no-punishment condition. This has the advantage that, in addition to across-subject comparisons, we can make

[2] Instructions are included in the long version of this paper which can be downloaded from our website (http://www.unizh.ch/iew/grp/fehr/index.html). The whole experiment was framed in neutral terms.

[3] Note that in the Partner-treatment the probability of being rematched with the same three people in the next period is 100 percent, whereas in the Stranger-treatment it is less than 0.05 percent. We also conducted experiments in which the probability of meeting the same subjects in future periods was exactly zero. Because of space constraints we do not present the results of these experiments. Contributions as well as punishment behavior in these perfect one-shot experiments are not significantly different from contributions and behavior in our Stranger-treatment. Hence, the Stranger-treatment represents a good approximation to perfect one-shot experiments.

within-subject comparisons of cooperation levels, which have much more statistical power. In Sessions 1–3 we implemented the Stranger-treatment, whereas in Sessions 4 and 5 we implemented the Partner-treatment. In Sessions 1 and 2 subjects first play ten periods in the punishment condition and then ten periods in the no-punishment condition. To test for spillover effects across conditions the no-punishment condition is conducted first in Session 3. In Session 4, which implemented the Partner-treatment, we start with the punishment condition, whereas Session 5 begins with the no-punishment condition.

B. *Payoffs*

In the following we first describe the payoffs in the treatments without punishment. In each period each of the n subjects in a group receives an endowment of y tokens. A subject can either keep these tokens for him- or herself or invest g_i tokens ($0 \leq g_i \leq y$) into a project. The decisions about g_i are made simultaneously. The monetary payoff for each subject i in the group is given by

$$(1) \quad \pi_i^1 = y - g_i + a \sum_{j=1}^{n} g_j, \qquad 0 < a < 1 < na$$

in each period, where a is the marginal per capita return from a contribution to the public good. The total payoff from the no-punishment condition is the sum of the period-payoffs, as given in (1), over all ten periods. Note that (1) implies that full free-riding ($g_i = 0$) is a dominant strategy in the stage game. This follows from $\partial \pi_i^1 / \partial g_i = -1 + a < 0$. However, the aggregate payoff $\sum_{i=1}^{n} \pi_i^1$ is maximized if each group member fully cooperates ($g_i = y$) because $\partial \sum_{i=1}^{n} \pi_i^1 / \partial g_i = -1 + na > 0$.

The major difference between the no-punishment and the punishment conditions is the addition of a second decision stage after the simultaneous contribution decision in each period. At the second stage, subjects are given the opportunity to simultaneously punish each other after they are informed about the individual contributions of the other group members. Group member j can punish group member i by assigning so-called punishment points p_j^i to i. For each punishment point assigned to i the first-stage payoff of i, π_i^1, is reduced by 10 percent. However, the first-stage payoff of subject i can never be reduced below zero. Therefore, the number of payoff-effective punishment points imposed on subject i, P^i, is given by $P^i = \min(\sum_{j \neq i} p_j^i, 10)$. The cost of punishment for subject i from punishing other subjects is given by $\sum_{j \neq i} c(p_i^j)$, where $c(p_i^j)$ is strictly increasing in p_i^j. The pecuniary payoff of subject i, π_i, from both stages of the punishment treatment can therefore be written as

$$(2) \quad \pi_i = \pi_i^1[1 - (1/10)P^i] - \sum_{j \neq i} c(p_i^j).$$

The total payoff from the punishment condition is the sum of the period-payoffs, as given in (2), over all ten periods.

C. *Parameters and Information Conditions*

The experiment is conducted in a computerized laboratory where subjects anonymously interact with each other.[4] No subject is ever informed about the identity of the other group members. In all treatment conditions the endowment is given by $y = 20$, groups are of size $n = 4$, the marginal payoff of the public good is fixed at $a = 0.4$, and the number of participants in a session is $N = 24$.[5] Table 2 shows the feasible punishment levels and the associated cost for the punisher. In each period subject i can assign up to ten punishment points p_i^j to each group member j, $j = 1, \ldots, 4$, $j \neq i$.

In all treatment conditions subjects are publicly informed that the condition lasts *exactly* for ten periods. When subjects play the first treatment condition in a session they do not know that a session consists of two conditions. After period ten of the first treatment condition in a session they are informed that there will be a "new experiment" and

[4] For conducting the experiments we used the experimental software "z-Tree" developed by Urs Fischbacher (1998).

[5] An exception is Session 4 where only $N = 16$ subjects showed up.

TABLE 2—PUNISHMENT LEVELS AND ASSOCIATED COSTS FOR THE PUNISHING SUBJECT

Punishment points p_i^j	0	1	2	3	4	5	6	7	8	9	10
Costs of punishment $c(p_i^j)$	0	1	2	4	6	9	12	16	20	25	30

that this experiment will again last exactly for ten periods. They are also informed that the experiment will then be definitely finished.

In the no-punishment conditions the payoff function (1) and the parameter values of y, n, N, and a are common knowledge. At the end of each period subjects in each group are informed about the total contribution $\sum g_j$ to the project in their group.

In the punishment conditions the payoff function (2) and Table 2, in addition to y, n, N, and a, are common knowledge. Furthermore, after the contribution stage subjects are also informed about the whole vector of individual contributions in their group. To prevent the possibility of individual reputation formation across periods in the Partner-treatment each subject's own contribution is always listed in the first column of his or her computer screen and the remaining three subjects' contributions are *randomly* listed in the second, third, or fourth column, respectively. Thus, subject i does not have the information to construct a link between individual contributions of subject j across periods. Therefore, subject j cannot develop a reputation for a particular individual contribution behavior. This design feature also rules out that i punishes j in period t for contribution decisions taken in period $t' < t$. Subjects are neither informed about the *individual* punishment activities of the other group members, nor do they know the *aggregate* punishment imposed on *other* group members. They know only their own punishment activities and the aggregate punishments imposed on them by the other group members.

II. Predictions

To have an unambiguous reference prediction it is useful to shortly state the implications of the standard approach to the public good games of Table 1. If the rationality and the selfishness of all subjects is common knowledge, and if subjects apply the backward induction logic, the equilibrium prediction with regard to g_i for each of the four cells in Table 1 is identical—in all four treatment conditions all subjects will contribute nothing to the public good in all periods. This is most transparent in the Stranger-treatment without punishment. This condition consists of a sequence of ten (almost pure) one-shot games. In each one-shot game the players' dominant strategy is to free ride fully. Applying the familiar backward induction argument to the Partner-treatment without punishment gives us the same prediction.

In the Stranger-treatment with punishment the situation is slightly more complicated because each one-shot game now consists of two stages. It is clear that a rational money maximizer will never punish at the second stage because this is costly for the player. Since rational players will recognize that nobody will punish at the second stage, the existence of the punishment stage does not change the behavioral incentives at the first stage relative to the Stranger-treatment without punishment. As a consequence, everybody will choose $g_i = 0$ at stage one. For the same reasons as in the Stranger-treatment rational subjects in the Partner-treatment with punishment will choose $g_i = 0$ and $p_i^j = 0$ for all j in the final period. By applying the familiar backward induction argument we thus arrive at the prediction that $g_i = 0$ and $p_i^j = 0$ for all j will be chosen by all subjects in all periods of the Partner-treatment with punishment.

There is already a lot of evidence for public good games like our no-punishment condition. For these games it is well known that cooperation strongly deteriorates over time and reaches rather low levels in the final period (John O. Ledyard, 1995). In a recent meta-study Fehr and Klaus M. Schmidt (1999) surveyed 12 different public good experiments without punishment where full free-riding is a dominant strategy in

the stage game. During the first periods of these experiments average and median contribution levels varied between 40 and 60 percent of the endowment. However, in the final period 73 percent of all individuals ($N = 1042$) chose $g_i = 0$ and many of the remaining players chose g_i close to zero. In view of these facts there can be little doubt that in the no-punishment condition subjects are not able to achieve stable cooperation. Therefore, a main objective of our experiment is to see whether subjects are capable of achieving *and* maintaining cooperation in the punishment condition.

In our view, the fact that at the beginning of the no-punishment condition one regularly observes relatively high cooperation rates, suggests that not all people are driven by pure self-interest. We conjecture that, in addition to purely selfish subjects, there is a nonnegligible number of subjects who are (i) conditionally cooperative and (ii) willing to engage in the costly punishment of free-riders. This conjecture is based on evidence from many other experimental games. Trust- or gift-exchange games (Fehr et al., 1993; Joyce Berg et al., 1995) indicate that many subjects are conditionally cooperative, that is, they are willing to cooperate to some extent if others cooperate, too. Bilateral ultimatum and contract enforcement games (e.g., Alvin E. Roth, 1995; Fehr et al., 1997) indicate that many subjects are willing to punish behavior that is perceived as unfair. In our public goods context fairness issues are likely to play a prominent role, too. We believe, in particular, that subjects strongly dislike being the "sucker," that is, being those who cooperate while other group members free ride. This aversion against being the "sucker" might well trigger a willingness to punish free-riders. In fact, recently developed theories of equity and fairness (e.g., Fehr and Schmidt, 1999) predict that free-riders will face credible punishment threats, which induces them to cooperate.

III. Experimental Results

In total, we have observations from 112 subjects. Each subject participated in only one of the five experimental sessions. All sessions were held in January and February 1996 at the University of Zurich (Switzerland). Subjects were students from many different fields (except economics). They were recruited via letters that were mailed to their private addresses. With this procedure we wanted to maximize the chances that subjects do not know each other. An experimental session lasted about two hours and subjects earned on average 41 Swiss francs (about US $32 at the time), including a show-up fee of 15 Swiss francs.

A. *The Impact of Punishment Opportunities in the Stranger-Treatment*

If subjects believe that in the presence of punishment opportunities free-riding faces no credible threat we should observe no differences in contributions across treatments. In sharp contrast to this prediction we can report the following result.

RESULT 1: *The existence of punishment opportunities causes a large rise in the average contribution level in the Stranger-treatment. On average, contribution rates amount to 58 percent of the endowment.*

Support for Result 1 is presented in Table 3. In columns 2 and 3 of Table 3 we report the mean contribution over all ten periods in the three sessions of the Stranger-treatment. The table reveals that in the punishment condition subjects contribute between two and four times more than in the no-punishment condition. A nonparametric Wilcoxon matched-pairs test shows that this difference in contributions is significant at all conventional significance levels ($p < 0.0001$). This result clearly refutes the hypothesis of the standard approach that punishment opportunities are behaviorally irrelevant at the contribution stage of the game.

Next we turn to the evolution of contributions over time. Remember that one of the most robust behavioral regularities in sequences of one-shot public good games, like our Stranger-treatment without punishment, is that contributions drop over time to very low levels. Our next result provides information as to whether punishment opportunities can prevent such a fall in contributions.

RESULT 2: *In the no-punishment condition of the Stranger-treatment average contributions converge close to full free-riding over time. In*

TABLE 3—MEAN CONTRIBUTIONS IN THE STRANGER-TREATMENT

	Mean contribution in all periods		Mean contribution in the final periods	
Sessions	Without punishment opportunity	With punishment opportunity	Without punishment opportunity	With punishment opportunity
1	2.7	10.9	1.3	9.8
	(5.2)	(6.1)	(4.3)	(6.8)
2	4.0	12.9	2.3	14.3
	(5.7)	(6.4)	(4.3)	(5.0)
3	4.5	10.7	2.0	13.1
	(6.0)	(4.9)	(3.8)	(4.0)
Mean	3.7	11.5	1.9	12.3
	(5.7)	(5.9)	(4.1)	(5.6)

Notes: Numbers in parentheses are standard deviations. Participants of Sessions 1 and 2 first played the treatment with punishment opportunities and then the one without such opportunities. Participants of Session 3 played in the reverse order.

contrast, in the punishment condition average contributions do not decrease or even increase over time.

Support for Result 2 comes from Table 3 and Figures 1A and 1B. Columns 4 and 5 of Table 3 show that, in each session, in the final period of the no-punishment condition average contributions vary between 1.3 and 2.3 tokens.[6] In contrast, in the punishment condition average contributions vary between 9.8 and 14.3 tokens in period ten. Thus, in the final period of the punishment condition the average contribution is between 6 and 7.5 times higher than in the no-punishment condition. Moreover, a comparison of column 3 with column 5 of Table 3 reveals that in the punishment condition the average contribution in period ten is higher or roughly the same as in all periods.

Figures 1A and 1B depict the evolution of average contributions over time in both conditions. Figure 1A shows the results of Sessions 1 and 2, in which subjects had to play the punishment condition first. Whereas the average contribution is stabilized around 12 tokens in the punishment condition, there is immediately a significant drop in contributions in period 11.[7] This decrease in the no-punishment condition continues until period 18 in which the average contribution stabilizes slightly below 2 tokens. Figure 1B shows the results of Session 3, in which subjects played the no-punishment condition first. In our view Figure 1B reveals an even more remarkable fact. Whereas average contributions in the no-punishment condition converge again toward 2 tokens they immediately jump upward in period 11 and *continue* to rise until they reach 13 tokens in period 20. This indicates that the existence of punishment opportunities triggers the effectiveness of forces that completely remove the drawing power of the equilibrium with complete free-riding. In view of this evidence it is difficult to escape the conclusion that any model which predicts full free riding is unambiguously rejected.

Results 1 and 2 deal only with average contributions. We are also interested, however, in the behavioral regularities at the individual level and how they are affected by the punishment opportunity. Result 3 summarizes the behavioral regularities in this regard.

RESULT 3: *In the Stranger-treatment with punishment no stable behavioral regularity*

[6] Note that in the following the term "final period" is always used to indicate the last period in a *given treatment condition* and not only period 20 in a given session. Thus, for example, in Figure 1A the tenth period is the final period of the punishment condition.

[7] The null hypothesis that average contributions are the same in period 10 and 11 can be rejected on the basis of a Wilcoxon signed-ranks test ($p = 0.0012$).

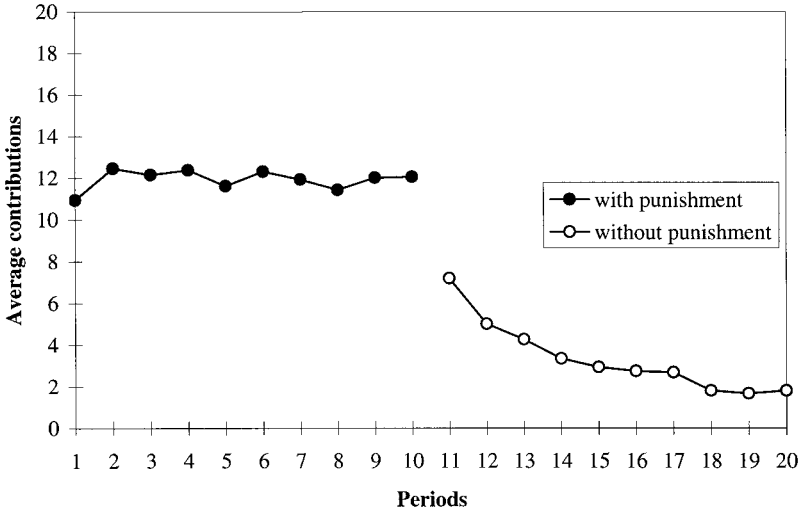

FIGURE 1A. AVERAGE CONTRIBUTIONS OVER TIME IN THE STRANGER-TREATMENT (SESSIONS 1 AND 2)

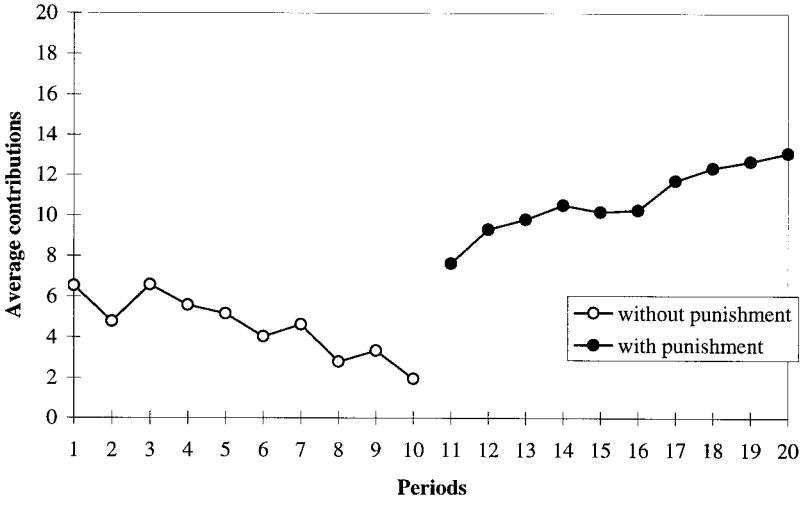

FIGURE 1B. AVERAGE CONTRIBUTIONS OVER TIME IN THE STRANGER-TREATMENT (SESSION 3)

regarding individual contributions emerges, whereas in the no-punishment condition full free-riding emerges as the focal individual action.

A first indication for the absence of a behavioral standard in the punishment condition is provided in Table 3. The table shows that the standard deviation of individual contributions is quite large in each session. Moreover, the standard deviation in the final period is roughly the same as in all periods together. This indicates that the variability of contributions does not decrease over time. The decisive evidence for Result 3, however, comes from Figure 2, which provides information about the relative frequency of individual choices in the final periods of both Stranger-treatments. In the no-punishment condition the overwhelming majority (75 percent) of subjects chose $g_i = 0$ in the final period. Thus, full free-riding clearly emerges as *the* behavioral regularity in this condition. In contrast, in the punishment condition individual choices are scattered over the whole strategy

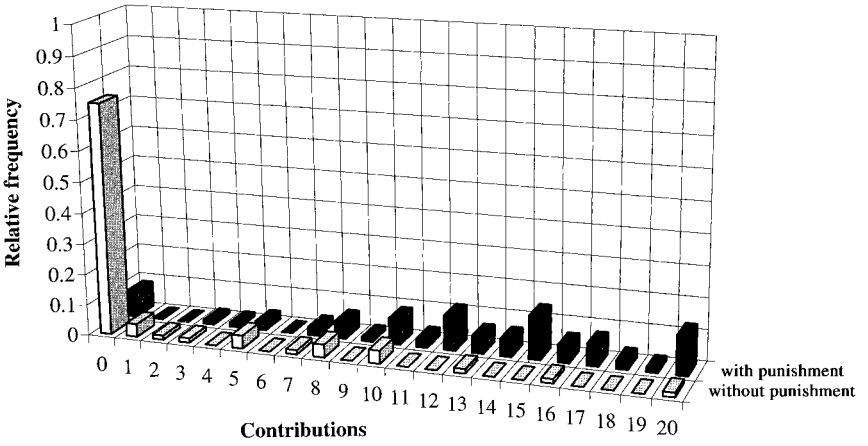

FIGURE 2. DISTRIBUTION OF CONTRIBUTIONS IN THE FINAL PERIODS OF THE STRANGER-TREATMENT WITH AND WITHOUT PUNISHMENT

space in the final period. Although the relative frequency of 12, 15, and 20 tokens is higher than that of other contribution levels, even the most frequent choice ($g_i = 15$) reaches a frequency of only 14 percent. Thus, subjects in the punishment condition were not able to coordinate on a specific contribution level different from $g_i = 0$.

B. *The Impact of Punishment Opportunities in the Partner-Treatment*

As in the Stranger-treatments our first result in the Partner-treatments relates to average contributions over all periods.

RESULT 4: *The existence of punishment opportunities also causes a large rise in the average contribution level in the Partner-treatment.*

Table 4 provides the relevant support for Result 4. A comparison of column 2 and column 3 shows that all ten groups have substantially higher average contributions in the punishment condition. Therefore, the difference is highly significant ($p = 0.0026$) according to a nonparametric Wilcoxon matched-pairs test with group averages as observations.

On average, subjects contribute between 1.5 times (group 2) and 4.3 times (group 9) more in the punishment condition. Thus, punishment opportunities are again highly effective in raising average contributions. With regard to the evolution of average contributions over time the data support the following result.

RESULT 5: *In the no-punishment condition of the Partner-treatment average contributions converge toward full free-riding, whereas in the punishment condition they increase and converge toward full cooperation.*

Again Table 4 provides a first indication. It shows that in the no-punishment condition the average contribution is only slightly above 3 tokens in the final period. In sharp contrast, the average contribution is above 18 tokens in the punishment condition. In five of the ten groups *all* subjects chose the maximum cooperation of 20 in the final period of the punishment condition. Further three groups exhibit average contributions of 19.3 or 19.5 tokens, respectively. A particularly remarkable fact represents the final period experience of group 9. Whereas *all* subjects chose full defection ($g_i = 0$) in the no-punishment condition *all* subjects chose full cooperation ($g_i = 20$) in the punishment condition.

Figures 3A and 3B show the evolution of average contributions over time. Irrespective of whether subjects play the punishment condition at the beginning or after the no-punishment condition, their average contributions in the final period are considerably higher than in the

TABLE 4—MEAN CONTRIBUTIONS IN THE PARTNER-TREATMENTS

	Mean contributions in all periods		Mean contributions in the final periods	
Groups	Without punishment opportunity	With punishment opportunity	Without punishment opportunity	With punishment opportunity
1	7.0	17.5	5.8	19.5
	(6.3)	(4.3)	(5.1)	(1.0)
2	10.6	16.4	1.0	19.3
	(8.5)	(5.2)	(1.4)	(1.5)
3	6.7	18.4	6.3	20.0
	(7.8)	(3.6)	(9.5)	(0.0)
4	5.1	12.1	1.3	13.5
	(6.3)	(7.1)	(2.5)	(8.5)
5	6.4	14.3	1.8	10.5
	(7.2)	(7.0)	(2.9)	(11.0)
6	7.9	19.0	3.5	20.0
	(5.7)	(2.8)	(5.7)	(0.0)
7	7.4	19.0	2.5	20.0
	(7.1)	(3.4)	(2.9)	(0.0)
8	10.0	17.2	5.0	20.0
	(6.6)	(4.3)	(6.0)	(0.0)
9	3.9	17.0	0.0	20.0
	(5.9)	(5.0)	(0.0)	(0.0)
10	10.0	19.0	5.0	19.5
	(6.6)	(2.1)	(8.0)	(1.0)
Mean	7.5	17.0	3.2	18.2
	(6.8)	(4.5)	(4.4)	(2.3)

Notes: Numbers in parentheses are standard deviations. Groups 1–4 (Session 4) first played the punishment condition and then the no-punishment condition. Groups 5–10 (Session 5) played in the reverse order.

first period of the punishment condition. The opposite is true in the no-punishment treatment. Moreover, at the switch points between the treatments there is a large gap in contributions in favor of the punishment condition. This indicates that the removal or the introduction of punishment opportunities immediately affects contribution behavior.[8] Thus, Table 4 and Figures 3A and 3B show that—in the Partner-treatment—punishment opportunities not only overturn the downward trend observed in dozens of no-punishment treatments; they also show that punishment opportunities render eight of ten groups capable of achieving almost *full* cooperation, although—according to the standard approach—*full* defection is the unique subgame perfect equilibrium.

A major purpose of the Partner-treatment with punishment is to enhance the possibilities for implicit coordination. We conjectured that this might enable subjects to converge toward a behavioral standard different from $g_i = 0$. Result 6 shows that this is indeed the case.

RESULT 6: *In the Partner-treatment with punishment, full cooperation emerges as the dominant behavioral standard for individual contributions, whereas in the absence of punishment opportunities full free-riding is the focal action.*

Evidence for Result 6 is given by Figure 4, which shows the relative frequency of indi-

[8] In Session 4 and in Session 5 average contributions in period 11 are significantly different from contributions in period 10 [Wilcoxon signed-ranks tests, $p = 0.05$ (Session 4) and $p = 0.027$ (Session 5)]. It is particularly remarkable that in Session 5 contributions in period 11 are even higher than in period 1 (Wilcoxon signed-ranks test, $p = 0.028$). *All* six groups of Session 5 contribute more in period 11 than in period 1.

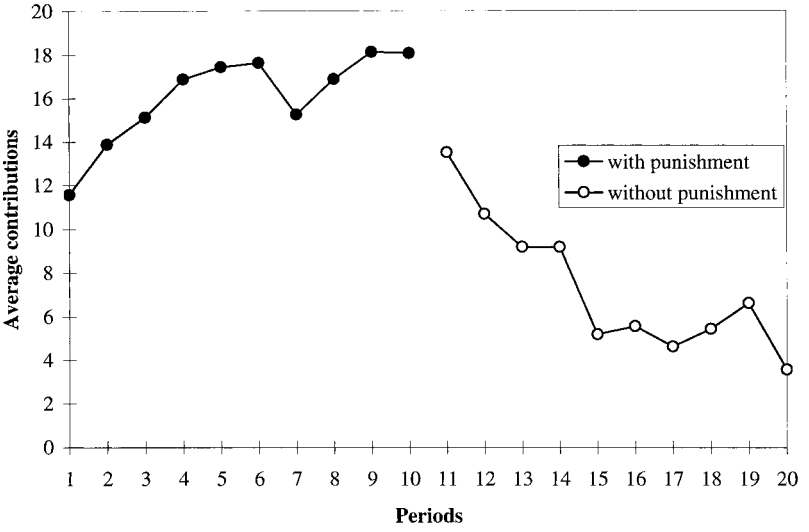

FIGURE 3A. AVERAGE CONTRIBUTIONS OVER TIME IN THE PARTNER-TREATMENT (SESSION 4)

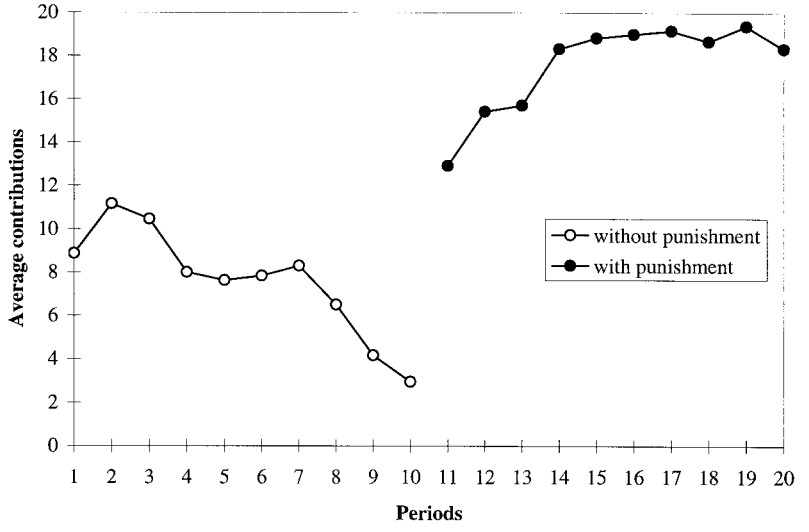

FIGURE 3B. AVERAGE CONTRIBUTIONS OVER TIME IN THE PARTNER-TREATMENT (SESSION 5)

vidual contributions in the final periods of the Partner-treatments. In the punishment condition 82.5 percent of the subjects contribute the whole endowment, whereas 53 percent of the *same* subjects free ride fully in the final period of the no-punishment condition. Moreover, in the no-punishment condition the majority of contributions is rather close to $g_i = 0$. The message of Figure 4 seems so unambiguous that it requires little further comment.

C. *Why Do Punishment Opportunities Raise Contributions?*

If there are indeed subjects who are willing to punish free-riding and if their existence is anticipated by at least some potential free-riders, we should observe that punishment opportunities have an *immediate* impact on contributions. Figures 1 and 3 show that this is indeed the case. After the introduction of punishment

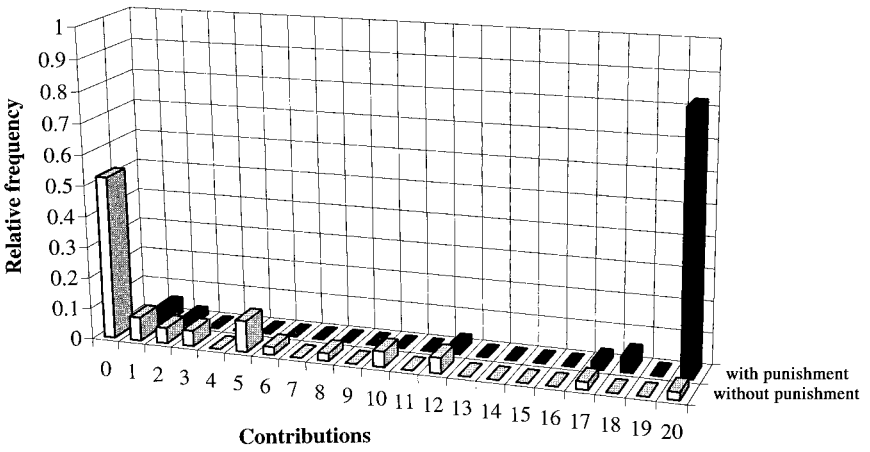

FIGURE 4. DISTRIBUTION OF CONTRIBUTIONS IN THE FINAL PERIODS OF THE PARTNER-TREATMENT WITH AND WITHOUT PUNISHMENT

opportunities in Session 3 (see Figure 1B) and Session 5 (see Figure 3B) there is an immediate increase in contributions. Moreover, after the removal of punishment opportunities in Sessions 1 and 2 (see Figure 1A) and Session 4 (see Figure 3A) contributions immediately drop to considerably lower levels. This suggests that potential free-riders are indeed disciplined in the punishment condition. A more detailed look at the regularities of actual punishments provides further support for this view.

RESULT 7: *In the Stranger- and the Partner-treatment a subject is more heavily punished the more his or her contribution falls below the average contribution of other group members. Contributions above the average are punished much less and do not elicit a systematic punishment response.*

Figure 5 and Table 5 provide evidence for Result 7. In Figure 5 we have depicted the average punishment levels as a function of negative and positive deviations from the others' average contribution in the group. For example, a subject in the Partner-treatment, who contributed between 14 and 20 tokens less than the average, received on average 6.8 punishment points from the other group members. The numbers above the bars indicate the relative frequency of observations in the different deviation intervals.

Figure 5 shows that in *both* treatments negative deviations from the average are strongly punished. Moreover, in the domain of negative deviations (i.e., in the three intervals below -2), the relation between punishment and deviations is clearly negatively sloped. The figure also indicates that there is a large drop in punishments if an individual's contribution is close to the average (i.e., in the interval $[-2, +2]$).[9] Finally, the figure suggests that positive deviations are much less punished and that the size of the positive deviation has only a weak impact on the punishment activities by other group members.[10]

[9] Figure 5 also provides further support for the emergence of a common behavioral standard for *individual* contributions in the Partner- but not in the Stranger-treatment. Note that 57 percent of all the individual contributions in the Partner-treatment are in the interval $[-2, +2]$, whereas only 26 percent are in this interval in the Stranger-treatment.

[10] One might ask why individuals with positive deviations get punished at all. According to a postexperimental questionnaire there are five potential reasons for this. (i) Random error. Since individuals can err on only one side at the punishment stage (i.e., rewarding others was not possible), each error shows up as a positive punishment. (ii) Subjects with very high individual contributions may view others' contributions as too low, even if they are above the average. (iii) Subjects may want to earn more than others (i.e., they punish, even if others cooperate, to achieve a *relative* advantage). (iv) Spiteful revenge. Free-riding subjects punish the cooperators because they expect to get punished by them. (v) Blind revenge. Subjects who get

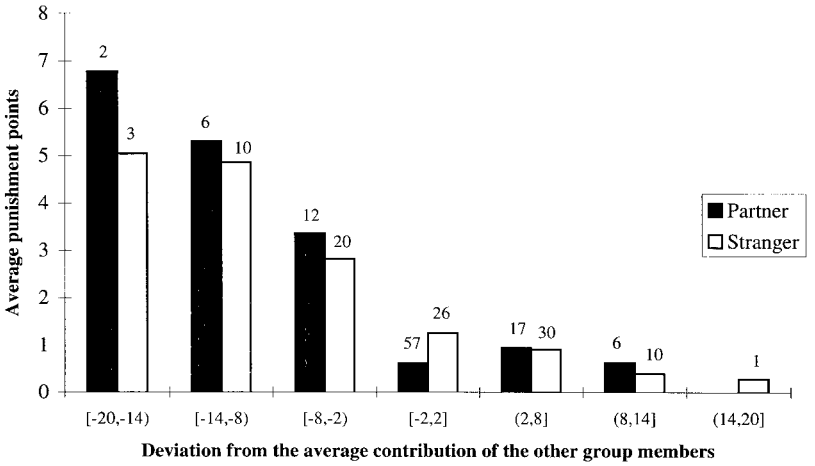

FIGURE 5. RECEIVED PUNISHMENT POINTS FOR DEVIATIONS FROM OTHERS' AVERAGE CONTRIBUTION

TABLE 5—DETERMINANTS OF GETTING PUNISHED: REGRESSION RESULTS

	Dependent variable: received punishment points	
Independent variables	Stranger-treatment	Partner-treatment
Constant	2.7363***	0.9881
	(0.0485)	(0.6797)
Others' average contribution	−0.0735***	−0.0108
	(0.0239)	(0.0457)
Absolute negative deviation	0.2428***	0.4168***
	(0.0325)	(0.0510)
Positive deviation	−0.0147	−0.0357
	(0.0264)	(0.0355)
	$N = 720$	$N = 400$
	$F[14, 705] = 39.0$***	$F[21, 378] = 41.3$***
	Adjusted $R^2 = 0.43$	Adjusted $R^2 = 0.68$
	DW = 1.96	DW = 1.89

Notes: Standard errors are in parentheses. * denotes significance at the 10-percent level, ** at the 5-percent level, and *** at the 1-percent level. To control for time and matching groups, the regression model also contains period dummies and dummies for matching groups (i.e., session dummies in the Stranger-treatment and dummies for each independent group in the Partner-treatment). Results are corrected for heteroskedasticity. Tobit estimations yield similar results.

To provide formal statistical evidence for Result 7 we also conducted a regression analysis of punishment behavior. Table 5 contains the model and the ordinary least-squares (OLS) regressions separately for the Stranger-treatment and the Partner-treatment. We also conducted Tobit regressions with the same variables. Yet, since they are similar to the OLS estimates we do not report them explicitly. The dependent variable is "received punishment points" of a subject and the independent variables comprise "others' average contribution" and the variables "positive deviation" and "absolute negative deviation," respectively. Figure 5 suggests

punished in $t - 1$ may assume that punishment was mainly exerted by the cooperators. By punishing cooperators in t they may take revenge. Note that by doing this they may punish the wrong target, because our design rules out the possibility of identifying individual contribution histories.

that positive and negative deviations from the others' average contribution elicit different punishment responses. These variables are therefore included as separate regressors. The variable "absolute negative deviation" is the absolute value of the actual deviation of a subject's contribution from the others' average in case that his or her own contribution is below the average. This variable is zero if the subject's own contribution is equal to or above the others' average. The variable "positive deviation" is constructed analogously. To model time effects, we included period dummies in the regression. The model also includes session dummies in the Stranger-treatment and group dummies in the Partner-treatment to control for fixed effects [see Manfred Königstein (1997)].

The results in Table 5 support the evidence from Figure 5. In both treatments the coefficient of the "absolute negative deviation" is positive and highly significant; thus, the more an individual's contribution falls short of the average the more that individual gets punished. In contrast, the size of the positive deviation has no significant impact on the size of the punishment. It is interesting that in the Partner-treatment it is *only* the negative deviation that affects punishment levels systematically, whereas the level of the others' average contribution has no significant impact. The low value and the insignificance of the coefficient on "others' average contribution" suggests that *only* deviations from the average were punished. This may be taken as evidence that in the Partner-treatment subjects quickly established a *common* group standard that did not change over time. If, instead, there would have been subjects who wanted to raise, say, the group standard, one should observe that a given negative deviation from the average is punished less the higher that average is. This is exactly what we observe in the Stranger-treatment in which the coefficient on "others' average contribution" is negative. The fact that there were subjects in the Stranger-treatment who wanted to raise the group standard is consistent with previous evidence which shows that subjects in the Stranger-treatment could *not* establish a common behavioral standard.

The pattern of punishment indicated by Figure 5 and Table 5 shows that free-riders can escape or at least reduce the received punishment substantially by increasing their contributions relative to the other group members. The response of subjects who actually were punished suggests that they understood this. In the Partner-treatment we observed 125 sanctions against subjects who contributed less than their endowment. In 89 percent of these cases the punished subject increased g_i immediately in the next period with an average increase of 4.6 tokens. In the Stranger-treatment we have 368 such cases. In 78 percent of these cases g_i increased in the next period by an average of 3.8 tokens. These numbers suggest that actual sanctions were rather effective in *immediately* changing the behavior of the sanctioned subjects. Subjects seemed to have had a clear understanding of why they were punished and how they should respond to the punishment.

D. *Payoff Consequences of Punishment*

A major effect of the punishment opportunity is that it reduces the payoff of those with a relatively high propensity to free ride. In the following we call those subjects "free-riders" who chose $g_i = 0$ in more than five periods of the no-punishment treatment. Twenty percent of subjects in the Partner-treatment and 53 percent in the Stranger-treatment obey this definition of a free-rider. In the Stranger-treatment with punishment opportunities the overall payoff of the free-riders is reduced by 24 percent relative to the no-punishment condition; in the Partner-treatment the payoff reduction is 16 percent. This payoff reduction is driven by two sources. First, free-riders are punished more heavily and second, they contribute more to the project in the punishment condition. On average, free riders raise their contributions between 10 and 12 tokens (i.e., by 50 to 60 percent of their endowment), relative to the no-punishment condition. However, there is also a force that works against the payoff reduction for free riders because the other subjects (the "nonfree-riders") also contribute more in the punishment condition. This limits the payoff reduction for the free-riders.

What are the aggregate payoff consequences of the punishment condition? To examine this question we compute the difference in the average group payoff between the punishment and the no-

punishment condition and divide this difference by the average group payoff of the no-punishment condition. This gives us the relative payoff gain of the punishment condition. Result 8 summarizes the evolution of the relative payoff gain for both the Partner- and the Stranger-treatment.

RESULT 8: *In both the Stranger- and the Partner-treatment the punishment opportunity initially causes a relative payoff loss. Yet, toward the end there is a relative payoff gain in both treatments. In particular, in the Stranger-treatment the relative payoff gain of the punishment condition is positive in the last two periods, whereas in the Partner-treatment it is positive from period 4 onward. In the final period the relative payoff gain is roughly 20 percent in the Partner-treatment and 10 percent in the Stranger-treatment.*

The temporal pattern of relative payoff gains results from two sources: (i) In the Partner-treatment, in particular, contributions are lower in the early periods of the punishment condition than during the later periods and this caused much more punishment activities in the early periods. (ii) Contributions gradually decline over time in the no-punishment condition. Taken together, Result 8 suggests that the presence of punishment opportunities eventually leads to pecuniary efficiency gains. To achieve these gains, however, it is necessary to establish the full credibility of the punishment threat by *actual* punishments.

IV. Conclusion

This paper provides evidence that spontaneous and uncoordinated punishment activities give rise to heavy punishment of free-riders. In the Stranger-treatment this punishment occurs, although it is costly and provides no future private benefits for the punishers. The more an individual negatively deviates from the contributions of the other group members, the heavier the punishment. Recently developed models of equity and reciprocity predict the widespread punishment of free-riders. Punishment is, however, clearly inconsistent with models of pure altruism or warm-glow altruism (e.g., James Andreoni, 1990) because an altruistic person never uses a costly option to reduce other subjects' payoffs. The apparent willingness to punish constitutes a credible threat for potential free riders and causes a large increase in cooperation levels: very high or even *full cooperation* can be achieved and maintained in the punishment condition, whereas the same subjects converge toward *full defection* in the no-punishment condition.

In our view punishment of free-riding also plays an important role in real life. It seems, for example, rather likely that—under team production—shirking workers elicit strong disapproval among their peers, and that strike-breaking workers face the spontaneous hostility of their striking colleagues. The enormous impact of the punishment opportunities on contributions in our experiment suggests that a neglect of the widespread willingness to punish free-riders faces the serious risk of making wrong predictions and, hence, giving wrong normative advice. Institutional and social structures that, theoretically, trigger the same behaviors in the absence of the willingness to punish may cause vastly different behaviors if the willingness to punish is taken into account.

REFERENCES

Andreoni, James. "Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving?" *Economic Journal*, June 1990, *100*(401), pp. 464–77.

Berg, Joyce; Dickhaut, John and McCabe, Kevin. "Trust Reciprocity and Social History." *Games and Economic Behavior*, July 1995, *10*(1), pp. 122–42.

Fehr, Ernst; Kirchsteiger, Georg and Riedl, Arno. "Does Fairness Prevent Market Clearing? An Experimental Investigation." *Quarterly Journal of Economics*, May 1993, pp. 437–60.

Fehr, Ernst; Gächter, Simon and Kirchsteiger, Georg. "Reciprocity as a Contract Enforcement Device—Experimental Evidence." *Econometrica*, July 1997, *65*(4), pp. 833–60.

Fehr, Ernst and Schmidt, Klaus M. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, August 1999, *114*(3), pp. 817–68.

Fischbacher, Urs. "z-Tree: Zurich Toolbox for Readymade Economic Experiments. Instructions for Experimenters." Mimeo, University of Zurich, 1998.

Francis, Hywel. "The Law, Oral Tradition and the Mining Community." *Journal of Law and Society*, Winter 1985, *12*(3), pp. 267–71.

Hirshleifer, David and Rasmusen, Eric. "Cooperation in a Repeated Prisoners' Dilemma with Ostracism." *Journal of Economic Behavior and Organization*, August 1989, *12*(1), pp. 87–106.

Königstein, Manfred. "Measuring Treatment Effects in Experimental Cross-Sectional Time Series." Mimeo, Humboldt-University, Berlin, 1997.

Ledyard, John O. "Public Goods: A Survey of Experimental Research," in John H. Kagel and Alvin E. Roth, eds., *Handbook of experimental economics*. Princeton: Princeton University Press, 1995, pp. 111–94.

Ostrom, Elinor; Walker, James and Gardner, Roy. "Covenants With and Without a Sword: Self-Governance is Possible." *American Political Science Review*, June 1992, *86*(2), pp. 404–17.

Roethlisberger, F. J. and Dickson, W. J. *Management and the worker: An account of a research program conducted by the Western Electric Company, Hawthorne Works*. Cambridge, MA: Harvard University Press, 1947.

Roth, Alvin E. "Bargaining Experiments," in John H. Kagel and Alvin E. Roth, eds., *Handbook of experimental economics*. Princeton: Princeton University Press, 1995, pp. 253–348.