Economics, values, and organization

Edited by

AVNER BEN-NER
University of Minnesota

LOUIS PUTTERMAN
Brown University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1998

Printed in the United States of America

Typeset in Times Roman

Library of Congress Cataloging-in-Publication Data

Economics, values, and organization / edited by Avner Ben-Ner,
Louis Putterman.
p. cm.
Revised versions of papers presented at the conference "Economics,
Values, and Organization," held at Yale University on April 19–21,
1996.
Includes index.
ISBN 0-521-58087-0 (hc)
1. Economics – Moral and ethical aspects – Congresses. 2. Social
norms – Congresses. 3. Values – Congresses. I. Ben-Ner, Avner.
II. Putterman, Louis G.
HB72.E28 1997
330–dc21 97–21733
CIP

A catalog record for this book is available from the British Library

ISBN 0 521 58087 0 hardback

CHAPTER 13

How effective are trust- and reciprocity-based incentives?

Ernst Fehr and Simon Gächter

1 Introduction

In modern economics people are conceptualized as being rational and selfish. On the basis of assumptions of rationality and selfishness economists have constructed a remarkable body of theoretical knowledge that allows precise predictions in a wide variety of circumstances. However, there remains the question whether the exclusive reliance on rationality and selfishness is capable of explaining people's actual behavior. We are convinced that there are conditions in which standard economic theory predicts and explains behavior quite well. On the basis of our research, we also believe that there are important and identifiable conditions in which these assumptions lead to empirically false predictions and may, therefore, generate wrong normative advice.

In this chapter we argue that there is an important class of conditions in which predictions that are based on purely selfish behavior are systematically violated. Such conditions regularly arise when it is impossible to enforce agreements completely. Standard economic theory predicts that agreements that are not fully enforceable will never be concluded because at least one of the involved parties will not meet its obligations. But, in turn, this will in general induce the other parties not to meet their obligations. Since everybody will anticipate that the parties will not meet their contractually specified duties it makes no sense to conclude the contract in the first instance. The fact that under conditions of incom-

We gratefully acknowledge financial support from the Swiss National Science Foundation (NO. 12-43590.95) and very helpful research assistance by Martin Brown, Armin Falk, Urs Fischbacher, Jean-Robert Tyran, and Paolo Vanini. We thank J. Baron, D. Messick, A. E. Roth, A. Ben-Ner, L. Putterman, and an anonymous referee for valuable comments.

pletely enforceable contracts many agreements cannot or will not be concluded gives rise to severe efficiency losses.[1]

This conclusion is radically changed if one assumes that people do not always fully exploit their opportunities to violate agreements at the expense of others. It may then become rational to enter agreements that are *not* fully enforceable. The major aim of this chapter is to show that under well controlled laboratory conditions this is the rule rather than the exception. Moreover, our results indicate that these deviations from the standard predictions are closely related to the notion of trust and reciprocity.

2 Trust and reciprocity under incompletely enforceable contracts

In the following we outline our main arguments in the context of a simple labor contracting example. We would like to stress, however, that our argument is more general and also applies to contractual relations beyond the employment relationship.

Suppose that a firm stipulates a contract which specifies a wage w and a required effort level $\hat{e}$. The firm has an enforcement technology that allows it to elicit *at most* an effort level of e_0 from selfish and rational workers. This means that, in case of $\hat{e} > e_0$, a worker who reduces effort below $\hat{e}$ will increase his net utility. Therefore, a rational profit maximizing firm who faces a selfish and rational worker cannot enforce $e > e_0$. How can the existence of reciprocity help the firm to elicit effort levels above e_0? To answer this question we have to define this term. Roughly speaking, reciprocity means that people respond to kind acts with kind behavior and if they are treated badly try to strike back. Moreover, they are willing to engage in such reciprocal behavior even if it is costly for them. To judge whether a certain behavior is kind or mean it is necessary to fix a reference standard. This standard may itself be affected by, e.g., the history of a relationship, the behavior in other similar relations, or the institutional environment.

For our present purposes it suffices to assume that the kindness of a certain contract offer $(w, \hat{e})$ is determined by the rent the worker receives from the firm. This rent is given by the utility from $(w, \hat{e})$ minus the opportunity costs of accepting $(w, \hat{e})$. If a worker is motivated by reciprocity considerations she will choose higher levels of e in response to

[1] By efficiency we mean the total gains that arise from trading between two parties. Throughout the chapter the notion of efficiency is only applied to the bilateral case. Thus, we do not consider cases in which two parties (e.g., two duopolists) strike an agreement at the expense of third parties (e.g., consumers).

higher rents offered. Thus, by paying sufficiently high rents firms may be capable of eliciting effort levels above e_0. From a psychological perspective there may be several reasons for workers' willingness to respond reciprocally. According to equity theory (Adams 1963, 1965) people try to equalize the ratio of perceived inputs (e.g., effort) and outputs (e.g., the rent) from a trade with the ratio that prevails in a relevant reference trade. Therefore, if the perceived output (rent) from a trade increases, people tend to raise their input (effort). A different explanation for reciprocal behavior is put forward by Rabin (1993), who argues that people's behavior vis-à-vis others is partly determined by their interpretation of the intentions that drive the behavior of others. They reward good intentions and punish bad intentions. Since the payment of a high rent is naturally interpreted as an action that is driven by a friendly intention workers may well reward this intention by a high effort level.

In the preceding example the firm offers a contract $(w, \hat{e})$. Once the worker has accepted the contract she chooses the actual effort level. Now suppose that after a worker's effort choice a firm has the option to reward or punish a worker. Both rewarding and punishing are costly for the firm. A rational and selfish firm will, therefore, never punish or reward. However, if the firm, i.e., the person acting on behalf of the firm, is motivated by reciprocity considerations, it may well reward $e \geqslant \hat{e} > e_0$ and punish $e < \hat{e}$. As before there may be several reasons for a firm's willingness to respond reciprocally. Firms may reward or punish for the reasons put forward by equity theory or by Rabin. There may, however, also be a third reason that has to do with the fact that contractual agreements usually have some normative force. If a worker accepts a contract $(w, \hat{e})$ with $\hat{e} > e_0$ she agrees to provide $e = \hat{e}$ even though only $e = e_0 < \hat{e}$ is in her selfish interest. The mere fact that she accepts the offer $(w, \hat{e})$ may be perceived by both parties as a kind of obligation to provide $\hat{e}$. It seems, moreover, likely that the perception of an obligation is the stronger the higher the rent implied by $(w, \hat{e})$. Since the violation of an obligation is likely to provoke moralistic aggression the firm may well be willing to punish the underprovision of effort. Analogously, the overfulfillment of an obligation may well trigger sympathy and, hence, a reward.

There are thus several potential reasons for reciprocal behavior. Our question in this chapter, however, is whether reciprocal behavior is indeed sufficiently strong to improve the enforcement of agreements significantly. During the last ten to fifteen years many experiments and several questionnaire studies which suggest that reciprocal behavior is quite common have been conducted. As indicated by Fehr, Kirchsteiger, and Riedl (1993, 1997), Fehr et al. (1997), and Berg, Dickhaut, and McCabe (1995), people often respond reciprocally if they receive a gift (rent). People's propensity to strike back if they are badly treated is suggested by the results of ultimatum games (see Roth et al. 1991; Güth and Tietz 1990; Roth 1995; Camerer and Thaler 1995). Questionnaire studies conducted by Agell and Lundberg (1995), Bewley (1995), Blinder and Choi (1990), and Kahneman, Knetsch, and Thaler (1986) also show that ordinary people as well as personnel managers believe that fairness, work morale, and reciprocity considerations are very important determinants of people's conduct and, in particular, of workers' effort behavior.

The evidence cited suggests that reciprocal behavior might also be relevant in the context of contract enforcement. The results of our experiments indicate that this is indeed the case. The observed regularities provide rather strong support for the relevance of reciprocal behavior to the enforcement of agreements. Even among anonymous strangers reciprocal interactions constitute a powerful means for the elicitation of effort levels above e_0. Subjects in the role of firms persistently demanded $\hat{e} > e_0$. Moreover these demands were associated with an appeal to reciprocity. Firms offered higher rents if they demanded more effort. This trust in reciprocal responses was justified in the sense that workers' *actual* effort is, on average, also positively related to the rent offered. If firms have *no* opportunity to punish or reward workers' effort choice, the underprovision of effort is quite common although much smaller than predicted by the standard approach. However, if firms can punish or reward ex post they are capable of substantially reducing the frequency of shirking. In fact, under these circumstances excess effort ($e > \hat{e}$) was more often observed than shirking ($e < \hat{e}$).

In the remainder of this chapter we describe the experimental design (section 3) and our results (section 4) in more detail. Section 5 concludes with a summary and raises some questions that have to be addressed by future work.

3 Experimental design

In this section we describe our experimental design.[2] To test for the impact of trust and reciprocity on workers' performance we developed two treatments: a two-stage treatment, in which only workers had a possibility to reciprocate, and a three-stage treatment, in which both workers and firms[3] had the opportunity to behave reciprocally.

[2] Instructions are available on request.

[3] In the experiment we did not use the possibly value-laden terms "workers" and "firms," but called them "buyers" and "sellers," respectively. The whole experiment was framed in goods-market terms. When we speak of "firms" and "workers" we mean of course subjects acting in the roles of firms and workers. For expositional simplicity we stick to the terms "firm" and worker.

3.1 *Common features in both treatments*

In total we conducted four experimental sessions (two sessions in both treatment conditions), which all took place at the University of Zurich in February 1996. Our subjects were students from the University of Zurich and the Federal Institute of Technology in Zurich. They were from different fields, yet no economists were among them. They were recruited via telephone calls to minimize the probability that subjects knew each other.[4] All of them received a show-up fee of 15 Swiss francs (about U.S.$12.5) plus their earnings in the experiment. During the experiment earnings where denoted in "points" and at the end of the experiment exchanged into Swiss francs with an exchange rate of 8 rappen (approximately 6 cents) per point.

The experiments were manually conducted and in all sessions we had eight workers and six firms, which were located in different rooms. In both treatments, at the very beginning of the experiment subjects were randomly allocated to rooms. After reading the instructions subjects had to answer a control questionnaire which tested their understanding of payoff calculations. We did not start an experiment before all subjects had correctly answered all questions. Both treatments involved a trial period which allowed subjects to become acquainted with the experimental procedures. To allow for learning and convergence we replicated the constituent stage games for an additional twelve periods. In all periods except the trial period subjects were paid according to the schedules detailed later.

3.2 *The two-stage treatment*

In both treatments the *first stage* was a posted-bid market. Firms posted an employment contract which consisted of a *wage* and a *desired effort level* $\hat{e}$.[5] Wages had to be integers between 0 and 100 and the available effort levels were between 0.1 and 1 with increments of 0.1 (see Table 1). In a given period each firm could offer only one employment contract; that is, each could employ only one worker. Firms made their decisions

[4] At the end of the experiments we asked subjects about their knowledge of other participants. It turned out that almost all of them had never met another participant before.

[5] In the experiments effort was framed as the "quality" of the experimental good and the term used for the desired effort level was "desired quality." We deliberately used goods-market terms instead of labor-market terms in the experimental instructions because we thought that labor-market terms might evoke more emotions and value-oriented behavior. Therefore, if trust and reciprocity show up in a goods-market framework we have a stronger result.

Table 1. *Effort levels and costs of effort*

effort *e*	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
costs of effort *c(e)*	0	1	2	4	6	8	10	12	15	18

privately but announced them publicly afterward. One experimenter in the firms' room wrote the offers on the blackboard whereas another experimenter listed the offers on a documentation sheet. After completion of all offers this documentation sheet was passed to the workers' room, where an experimenter wrote the offers on the blackboard in a randomly assigned order. This procedure was implemented to inhibit identification of firms. After all offers had been written on the blackboard, workers could choose among them in a randomly determined order. A worker could only accept one offer. As there were eight workers and six firms, we always had an excess supply of two workers. Therefore, in any period at least two workers could not or did not accept a contract.[6]

After the acceptance of offers the first stage was completed and the *second stage* started. Workers who had accepted an offer had to determine their *actual effort level.* A worker's choice of an effort level was associated with costs for the worker as indicated in Table 1. Of course, firms had the same effort levels available when choosing their desired effort, and they were informed about the associated costs for the worker. Table 1 was common knowledge. Workers privately determined their effort levels by inserting them into a decision sheet which was distributed to them at the beginning of the experiment. An experimenter collected the effort decisions and passed them to the firms' room. No worker was informed of the effort decision of fellow workers. Except the experimenter, only the firm with whom a worker concluded an employment contract was informed of the effort decision of the worker. However, firms were not informed of the identity of "their" worker.

Our second stage reflects a basic feature of most employment relationships, namely, the incompleteness of the labor contract. Workers in

[6] If, for example, the first six workers accepted an offer there were no available offers for the last two workers and, hence, they could not accept an offer. A worker was of course free to reject any available offer.

the real world almost always have considerable discretion in determining their actual effort.[7] Therefore, in our design firms could only stipulate a *desired* effort level without being able to enforce it (e.g., with the help of courts). The only effort level which was enforceable was the minimum effort level of 0.1.

After firms had been informed about the effort decision of their worker the second stage was completed and payoffs could be calculated. A worker's payoff at the end of the second stage was given by

$$u = w - c(e)$$

where w denotes the accepted wage and $c(e)$ the costs of the worker's actual effort. If a worker did not trade, she earned nothing. A firm's payoff was given by

$$\pi = ve - w$$

where v denotes a firm's redemption value. In all sessions v was equal to 100.

It is immediately apparent from this payoff function and Table 1 that firms may suffer a (severe) loss if they offer a wage which is higher than 10 and if they get an actual effort level of 0.1.[8] We therefore gave subjects – in addition to their show-up fee – an endowment of 112.5 points. Hence, given our exchange rate, subjects had – together with their show-up fee – a total start-up endowment of 300 points. Subjects were told that if their total earnings (including the 300 points) became negative they would have to leave the experiment.[9] To prevent a loss of control over subjects' preferences (possibly because of feelings of envy) both workers and firms received 112.5 points as a start-up endowment. Because this is a lump-sum payment, marginal incentives are not affected.

Our research concerns the potential of trust and reciprocity in enhancing enforceable effort levels. As reciprocity means kindness to those who are kind to you and meanness to those who are mean to you, it requires the possibility of judging the generosity of an employment offer. Therefore, payoff functions were common knowledge.

[7] See for example the interview studies by Levine (1993) and Bewley (1995). In both studies the interviewed personnel managers pointed out the importance of "worker morale," which is only partly under a firm's control.

[8] Of course firms could avoid losses with certainty by offering a wage below 10.

[9] In fact, in our experiments not a single subject had a loss which would have forced him or her to leave the experiment.

3.3 *The three-stage treatment*

Notice that in our two-stage design it is basically the workers who have the opportunity to respond reciprocally to a firm's offer. However, most actual labor relations are long-term relationships and firms also have opportunities to react to a worker's effort decision. For example, firms usually have many possibilities to influence a worker's utility of a given job, e.g., promotion policies, access to fringe benefits, and social sanctioning. To prevent the problems of modeling a repeated game, we introduced a *third stage* into our design, in which firms had the opportunity to punish or to reward their worker at some cost. In particular, after having learned a worker's effort decision, a firm could choose a punishment/reward variable $p \in [-1, 1]$. Punishing ($p < 0$) or rewarding ($p > 0$) was costly for a firm. Only in the case of $p = 0$, i.e., when the firm neither rewards nor punishes, would no costs arise. A firm could only punish its worker if she had shirked. Likewise, rewarding was only possible if the worker chose at least the desired effort level. A punishment effected a reduction of a worker's gain at the end of the second stage by $25p$, whereas a reward gave the worker an additional payment of $25p$. The feasible punishment/reward levels as well as firm's costs are depicted in Table 2, which was common knowledge. Workers were privately informed about the punishment/reward choice of "their" firms. The payoffs at the end of the third stage were given by

$$u = w - c(e) + 25p$$

for a worker and by

$$\pi = 100e - w - k(p)$$

for a firm, where $k(p)$ denotes the firm's cost of punishing or rewarding.[10] To get some insight into workers' expectation formation when determining their actual effort at the second stage we asked workers in our three-stage treatment to insert into their decision sheet the *punishment/reward level they expect from their firm, given their effort decision.* Firms were not informed of the expected punishment/reward level.

In all other respects the first and the second stages were identical to our two-stage design. The Appendix summarizes our experimental procedure.

[10] In the experiment we called punishment and reward variables "negative" and "positive factors," respectively. Costs were framed as "costs of the factor." The payment to the worker was called "additional payment."

Table 2(a). *Punishment variable and costs of punishment for the firm*

punishment variable	0	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9	-1
costs for firm *k(p)*	0	1	2	3	4	5	6	7	8	9	10

Note: The punishment levels -0.1 to -1 were only available to a firm if a worker had shirked, i.e., if the actual effort level fell short of the desired effort level *ê*.

Table 2(b). *Reward variable and costs of rewarding for the firm*

reward variable	0	+0.1	+0.2	+0.3	+0.4	+0.5	+0.6	+0.7	+0.8	+0.9	+1,0
costs for firm *k(p)*	0	1	2	3	4	5	6	7	8	9	10

Note: The reward levels 0.1 to 1 were only available to a firm if a worker had chosen an actual effort level at least as high as the desired effort level *ê*.

4 Experimental results

In each of our two treatments there were in total 144 possible trades, of which 141 trades were realized in both treatments. The two-stage experiments lasted about 2.5 hours and the three-stage experiments about 3 hours. On average, a subject earned $48 in a two-stage and $58 in a three-stage session. In section 4.1 we describe our main results, which are followed by our results on effort elicitation in the two-stage treatment (section 4.2) and the three-stage treatment (section 4.3). Section 4.4 presents results on the distribution of incomes in our two treatments.

4.1 *Main results*

In our treatments the game-theoretic predictions are straightforward: In the two-stage treatment a worker will in each period choose the lowest possible effort level ($e = 0.1$), because higher effort levels are increasingly costly (see Table 1). Firms anticipate this effort choice and make a job offer with a desired effort of 0.1 and a zero wage. This prediction is not changed in the three-stage design, because the third stage only adds dominated strategies. To punish or to reward is a "noncredible threat," which, therefore, does not alter decisions at the first and the second stage of the game.

In the following we confront these predictions with the results of our experiments. The first result concerns firms' *desired* effort levels, whereas the second result documents workers' *actual* effort decision.

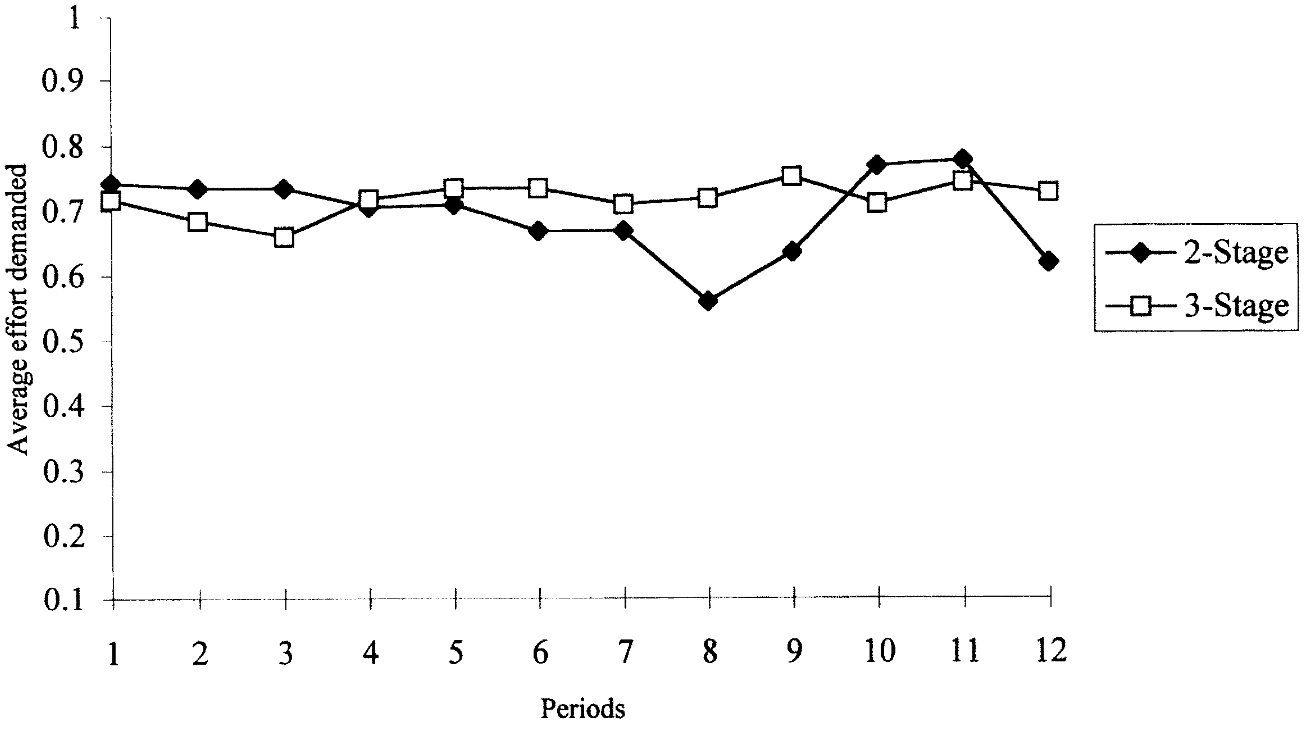

Figure 1. Firms' average desired effort in the two- and three-stage treatment

Result 1: In the two- as well as in the three-stage treatment firms demand effort levels that are not incentive compatible.

Figure 1 shows that desired effort levels are clearly above the theoretically predicted level of 0.1: On average firms demand an effort level of 0.70 in the two-stage treatment and of 0.72 in the three-stage treatment, with standard deviations of 0.25 and 0.13, respectively.[11] Moreover, in both treatments, firms' desired effort levels do not change very much over time. If anything, their desired effort level in the two-stage treatment is a bit more volatile than in the three-stage treatment. In addition, Figure 1 reveals that in both treatments there is clearly no convergence to the theoretically predicted level of 0.1. How successful were firms in the elicitation of actual effort levels? Our second main result shows that firms indeed succeeded in eliciting higher than incentive-compatible effort levels.

Result 2: In the two- and in the three-stage treatment firms are capable of eliciting non-incentive-compatible effort levels.

In the two-stage treatment firms were able to elicit an average actual effort level of 0.44 (with a standard deviation of 0.27), whereas in the three-stage treatment they could enforce an even higher mean actual effort level of 0.63 (standard deviation of 0.29). Figure 2 shows that actual effort levels as well do not converge to 0.1 even though in both treatments actual efforts are more volatile than desired effort levels.

Our third main result concerns the effectiveness of the third stage, which, from a game-theoretic point of view, should have no behavioral influence.

Result 3: Actual effort is on average higher in the three-stage treatment compared to the two-stage treatment.

This result confirms that the third stage is indeed behaviorally relevant. In the three-stage design firms can enforce effort levels which are on average 0.21 unit higher than in the two-stage treatment. This difference is significant at any conventional level (Mann-Whitney U test; trades as observations). Moreover, in the third stage shirking is considerably reduced compared to that in the second stage, as the following result shows.

[11] A Mann-Whitney U-test shows that the null hypothesis of equal means cannot be rejected at the 5 percent significance level. For a discussion of all nonparametric tests used in this chapter see Siegel and Castellan (1988).

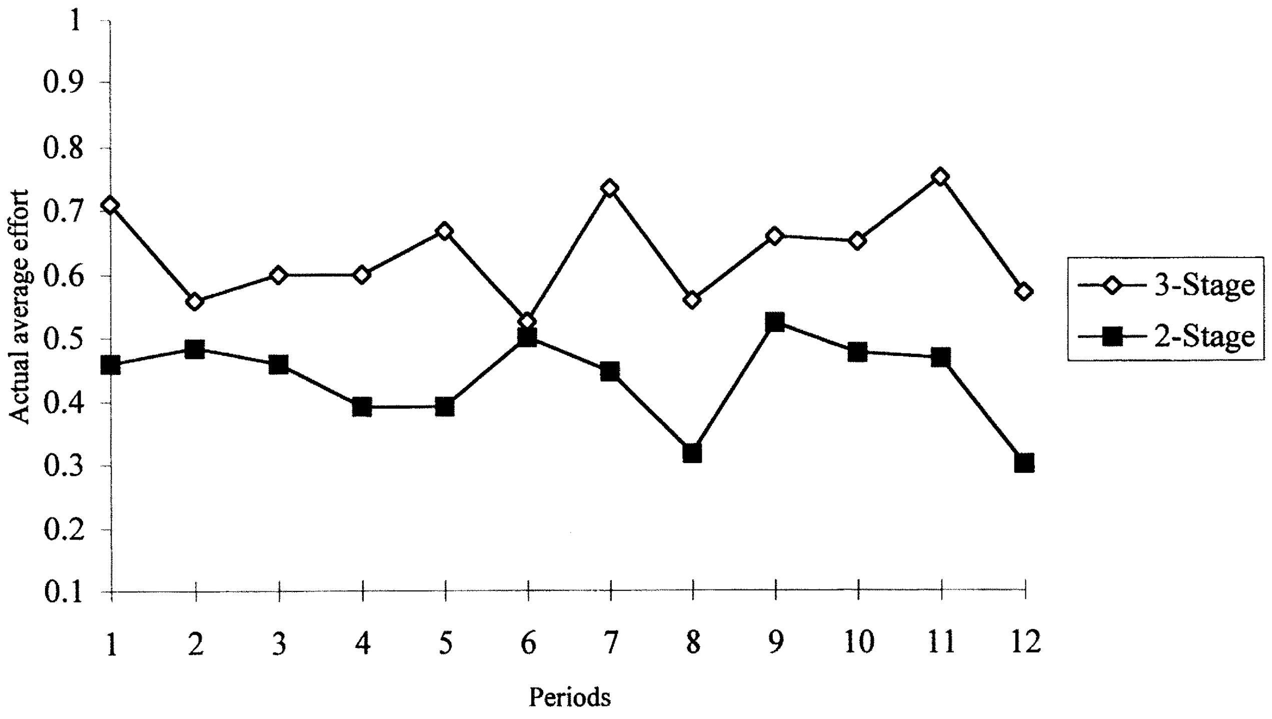

Figure 2. Workers' average actual effort in the two- and three-stage treatment

Table 3. *Effort behavior in the two- and three-stage treatment*

treatment	**No. trades**	**shirking** $e < \hat{e}$		**no shirking** $e = \hat{e}$	**excess effort** $e > \hat{e}$	
		% of trades with $e < \hat{e}$	average amount of $(\hat{e}-e)$	% of trades with $e = \hat{e}$	% of trades with $e > \hat{e}$	average amount of $(e-\hat{e})$
2-stage	141	82.98	0.31	14.18	2.84	0.18
3-stage	141	26.24	0.54	36.17	36.88	0.16

Result 4: The existence of a third stage reduces shirking significantly and generates excess effort in many trades.

This result follows from Result 3 and the fact that desired effort levels do not differ on average across treatments. To gain more insights into the effects of stage three we analyzed workers' effort behavior in more detail. Logically effort decisions fall into three categories: A worker has shirked (i.e., the actual effort e fell short of the desired effort $\hat{e}$ [$e < \hat{e}$]), she has just delivered the desired effort ($e = \hat{e}$), or she has delivered a higher effort than asked for ($e > \hat{e}$). In Table 3 we discuss shirking behavior separately for each of these categories.

Table 3 reinforces Result 3. Shirking rates are considerably reduced in the three-stage treatment compared to the two-stage treatment. In the two-stage treatment we observed some shirking in 82.96 percent of all trades (with an average shirking of 0.31).[12] Excess effort was very rarely observed: Workers delivered an average excess effort of 0.18 (2.84 percent) in only 4 trades. No shirking was a bit more common: It occurred in 14.18 percent of all trades.

This picture is completely different in the three-stage treatment. Shirking occurred in considerably fewer trades than in the two-stage treatment. Workers only shirked in 26.24 percent of trades. Yet, if they shirked, the amount of shirking was larger than in the two-stage treatment. No shirking happened in 36.17 percent of the trades, and, what is particularly interesting, in 36.88 percent of trades we observed excess effort. In view of the already rather high levels of the desired effort this large number of excess effort choices is remarkable.

[12] Shirking behavior in the two-stage treatment may have been influenced by a subject's normative considerations of contract fulfillment. The fact that somebody accepted a contract renders it more difficult to behave opportunistically ex post. For psychological explanations of self-commitment see Cialdini (1993, chap. 3).

4.2 *Effort elicitation in the two-stage treatment*

Given our main results the question of how firms were able to enforce non-incentive-compatible effort levels arises. On the basis of evidence from previous research (Fehr, Kirchsteiger, and Riedl 1993; Fehr et al. 1997; Fehr, Gächter, and Kirchsteiger 1997) we believe that the impulse to behave reciprocally constitutes an important part of the answer to this question. For most human beings reciprocal responses are a "natural" part of their behavioral dispositions. To be kind (mean) to those who have been kind (mean) to us seems to be a natural behavioral response. Moreover, the higher the level of kindness that somebody experiences, the stronger will, in general, be the impulse to reciprocate.

To operationalize the concept of reciprocity it is of course necessary to define how kind or mean an action is. In our context the rent $\hat{u} = w - c(\hat{e})$ that is offered to the workers can be taken as an indicator of "generosity." The higher the rent $\hat{u}$, the kinder is the firm to the worker. The kindness of the worker is in turn indicated by the actual effort level. In our two-stage treatment reciprocity, therefore, means that workers choose higher effort levels if firms offer them higher rents. In addition, an appeal to workers' reciprocity requires that firms offer higher rents if they want to induce higher effort levels, i.e., if they desire a higher $\hat{e}$. This subsection investigates whether these predictions are met by the data. Our first result in this respect is related to firms' behavior.

Result 5: Firms appeal to workers' reciprocity; i.e., the rent offered is positively related to the desired effort level in the two-stage treatment.

Figure 3 shows the average offered rent for a given level of effort demanded in the two-stage treatment. The higher the desired effort level, the higher was on average the offered rent. This is also confirmed by the nonparametric Spearman rank-order test (correlation coefficient $\rho = 0.73$, $p < 0.001$).

Notice that this result means that firms overcompensate their workers. When they demand higher levels of $\hat{e}$ they increase the wage by more than the cost increase for the worker. This result directly contradicts the prediction of the theory of compensating wage differentials. Our next result confirms that the appeal to reciprocity was indeed a successful strategy.

Result 6: Workers respond reciprocally; i.e., actual effort is positively related to the offered rent in the two-stage treatment.

Figure 4 documents workers' average actual effort as a function of firms' offered rents (depicted in intervals of 5). It clearly shows a positive

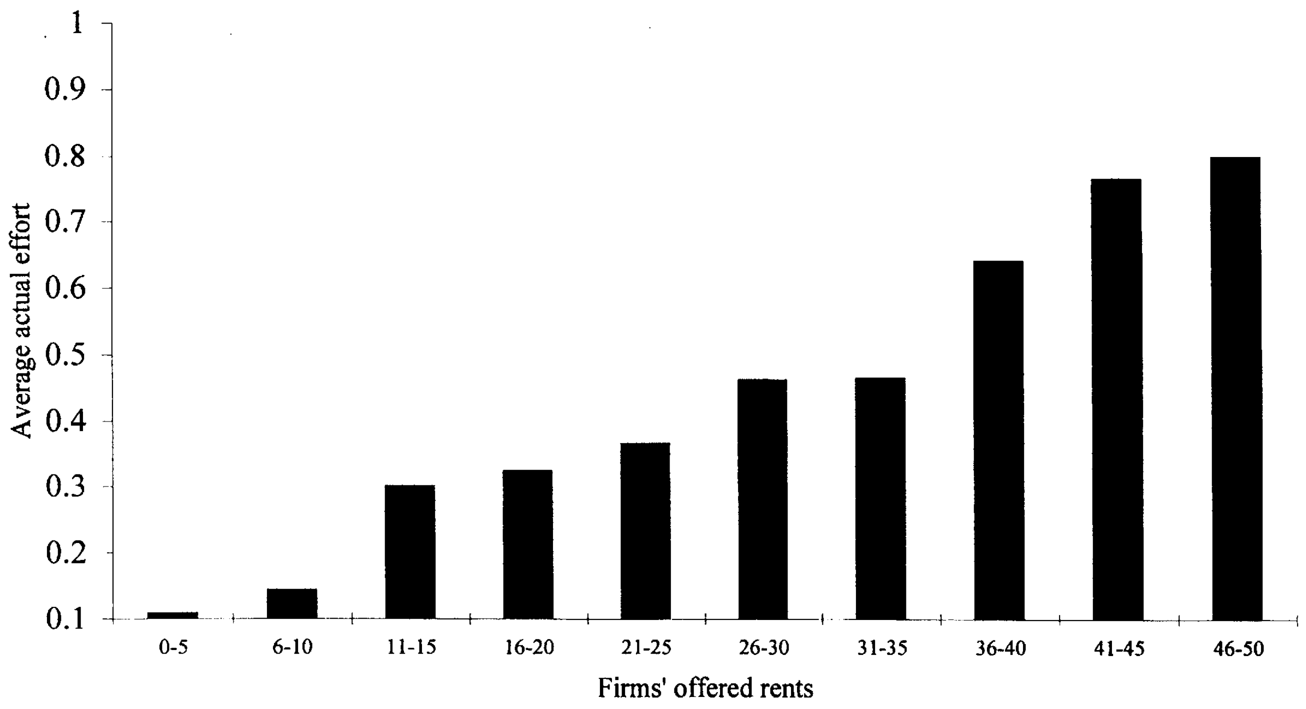

Figure 3. Workers' average actual effort given firms' offered rents

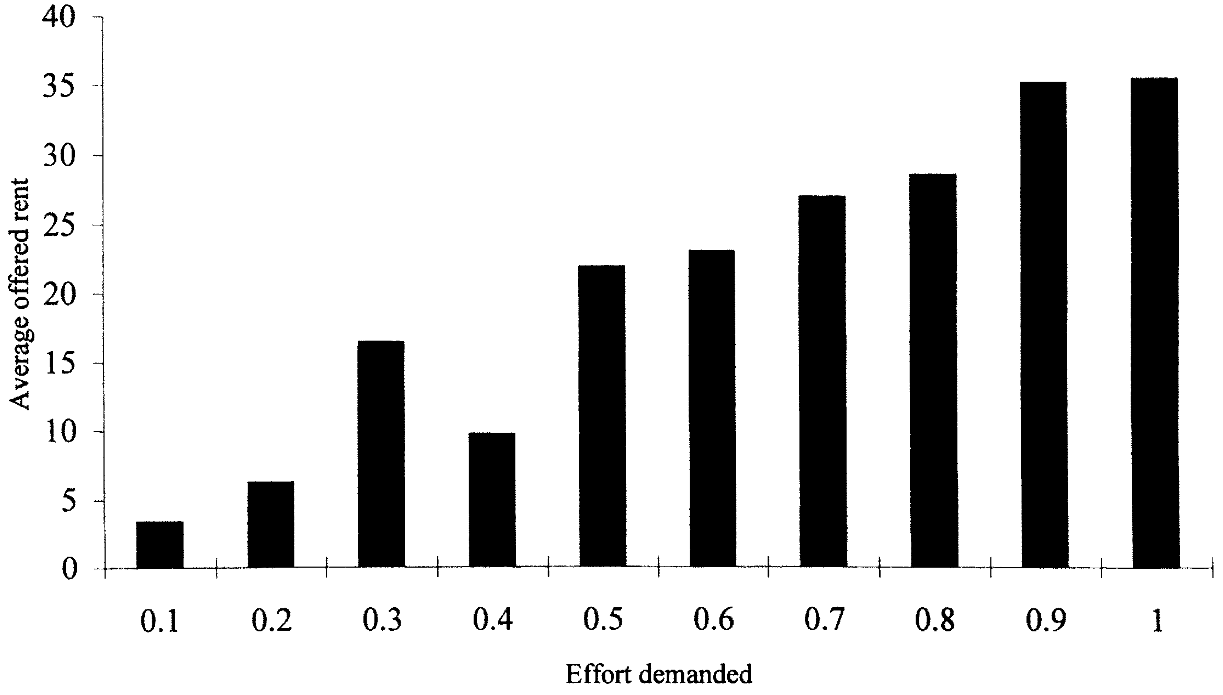

Figure 4. Firms' average offered rents per effort demanded

Table 4(a). *Firms' punishment/reward decision at stage three, given workers' effort decision*

actual punishment/reward	shirking $e < \hat{e}$ 37 trades	no shirking $e = \hat{e}$ 51 trades	excess effort $e > \hat{e}$ 53 trades
$p < 0$	67.6 % (-0.71)	not possible	not possible
$p = 0$	32.4 %	58.8 %	29.7 %
$p > 0$	not possible	41.2 % (0.45)	70.3 % (0.7)

Note: Number in parentheses indicates average level of p.

relationship between offered rents and workers' actual effort. This is also confirmed by a Spearman rank correlation using trades as observations ($\rho = 0.67$, p-value < 0.001).

We now turn to the three-stage treatment, in which firms also have an opportunity to behave reciprocally, because they can punish or reward their worker for his or her effort decision.

4.3 *Effort elicitation in the three-stage treatment*

We have already seen in Results 3 and 4 that the third stage was quite effective with respect to the elicitation of effort levels. We now analyze the behavior of firms at stage three. Remember that from a game-theoretic point of view rational firms will never punish or reward, because this is costly for them; see Tables 2(a) and 2(b), respectively. The actual behavior of firms in our experiment, however, is succinctly summarized by our Result 7.

Result 7: Firms behave reciprocally at stage three; i.e., they reward excess effort and punish shirking.

We substantiate this result in two steps. In Table 4(a) we summarize firms' actual punishment/reward behavior given workers' effort decision at the second stage. We then discuss some regularities of firm behavior using nonparametric correlation analysis. Table 4(a) shows firms' punishment/reward behavior given workers' effort choices at stage two. In 37 trades workers have shirked, whereas in 51 trades workers have just delivered the desired effort. In 53 trades workers have delivered a higher effort than asked for.

Only in a case in which a worker has actually shirked did a firm have the possibility of punishing the worker by choosing $p < 0$; see Table 2(a). Firms punished shirking workers in 25 trades (67.6 percent) and chose an average punishment level of –0.71. In case workers did not shirk or even exerted an excess effort, firms had the possibility of rewarding the workers by choosing $p > 0$. In 70.3 percent of the trades in which workers delivered excess effort, firms actually rewarded their workers, by choosing on average $p = 0.7$. In both the shirking and excess effort cases the selfishly rational choice of $p = 0$ was taken in less than a third of these cases. Firms also rewarded workers (with an average of $p = 0.45$) in 42.5 percent of the no-shirking trades.[13]

Table 4(a) clearly substantiates the importance of firms' reciprocal behavior. Firms exhibited *positive reciprocity* when workers chose excess effort and *negative reciprocity* when workers shirked. Chi-square tests [which test the null hypotheses that $p <$ (>) 0 is equally as likely as that $p = 0$ in the cases of shirking and excess effort, respectively] confirm that reciprocal behavior is more likely than the money maximizing choice of $p = 0$. In both cases the null hypothesis has to be rejected at conventional significance levels in favor of the alternative hypotheses, $p < 0$ and $p > 0$, respectively.[14]

As Result 6 shows, *workers'* behavior in the two-stage treatment shows clear regularity. Not only do they choose nonminimal effort levels in response to positive rents, but there is also a strong positive correlation between the actual effort and the rent level. Although Table 4(a) demonstrates the importance of firms' reciprocity, one may ask whether a similar regularity to that in workers' effort decisions shows up in firms' punishment/reward choice. In particular, we investigated whether

(i) in case of shirking there is a negative relationship between the degree of shirking $(\hat{e} - e)$ and the level of punishment $(p \leqslant 0)$, and whether

(ii) in case of excess effort there is a positive relation between the degree of excess effort $(e - \hat{e})$ and the level of reward $(p \geqslant 0)$.

According to Spearman rank-order tests using trades as observations the answer is clearly positive in the case of shirking. Firms punish more

[13] An interpretation of this behavior is that firms have rewarded workers for not exploiting their second mover advantage.

[14] In the case of no shirking the null hypothesis of an equal number of observations of $p =$ 0 and $p > 0$, has to be rejected in favor of the alternative $p = 0$, which means that in the case of no shirking (but no excess effort) it is more likely to receive no reward.

Table 4(b). *Workers' expectation formation: Do they anticipate firms' reciprocity?*

expected punishment/reward	**shirking** $e < \hat{e}$ 37 trades	**no shirking** $e = \hat{e}$ 51 trades	**excess effort** $e > \hat{e}$ 53 trades
$p^e < 0$	54.1 % (-0.42)	not possible	not possible
$p^e = 0$	45.9 %	37.3 %	1.9 %
$p^e > 0$	not possible	62.7 % (0.36)	98.1 % (0.64)

Note: Number in parentheses indicates average level of p^e.

the higher the degree of shirking ($\rho = -0.49$, p-value < 0.001). In the case of excess effort there is no such monotonic relation ($\rho = -0.07$, p-value $= 0.339$). However, if we restrict the analysis to cases in which firms actually punish ($p < 0$) or reward ($p > 0$) firms clearly punish more the higher the degree of shirking and reward more the higher the excess effort ($\rho = -0.59$, p-value < 0.001, and $\rho = 0.41$, p-value < 0.001, respectively). We take this as further evidence for firms' reciprocal behavior at stage three.

Next we ask whether workers' expectations concerning firms' punishment/reward behavior are "rational" given firms' actual behavior at stage three. Remember that we asked workers to report their expected punishment/reward after making their effort decisions (see section 3.3). Table 4(b) summarizes workers' expectations concerning firms' reciprocity. In those cases in which workers delivered an excess effort, they expected in 98.1 percent of the trades that they would be rewarded. In those trades they expected an average reward of 0.64. Shirking workers expected only in 54.1 percent of the trades to be punished (with an expected level of –0.42), whereas in 45.9 percent of the trades they believed that their firm would choose $p = 0$. According to a chi-square test one cannot reject the null hypothesis that shirking workers regarded punishment and no punishment as equally likely. Interestingly, however, workers who delivered exactly the desired effort expected in almost two-thirds (62.7 percent) of these trades that they would be rewarded (on average with 0.36). A chi-square test reveals that reward is indeed regarded as more likely than no reward in those trades in which workers delivered $e = \hat{e}$.

To summarize, workers *correctly expected* that they would be rewarded in the case of excess effort. They were, however, too optimistic in the other cases. When they exactly met the effort requirement they too often expected a reward, and when they shirked they underestimated the actual number of trades in which they would be punished. One can also ask how "rational" workers' expectations were concerning the *level* of punishment or reward, respectively. A comparison of the excess effort columns of Table 4(a) and Table 4(b) reveals that on average workers expected a reward of 0.64. Actually they received an average reward of 0.55.[15] To test whether expected and actual reward levels are equal, we conducted a Mann-Whitney test. In this test we compared levels of *expected* reward with levels of *actual* reward over all periods. The results of these tests do not allow us to reject the null hypothesis that expected and actual reward levels are equal (p-value = 0.5644). In other words, workers not only correctly expected rewards but also anticipated actual *levels* of firms' reward choices.[16]

In the case of shirking, however, the Mann-Whitney test reveals that shirking workers underestimated not only firms' propensity to punish shirking behavior at all, but also the level of actual punishment. In particular, the null hypothesis that expected and actual levels of punishment are equal has to be rejected in favor of the alternative hypothesis that actual punishment is more severe than expected punishment (p-value = 0.0109).

We summarize our findings about workers' expectations in the following.

Result 8: Workers correctly anticipate firms' rewards but underestimate firms' willingness to punish shirking.

4.4 *Do "noncredible" threats affect the distribution of income?*

The opportunity for firms to punish or reward their workers at stage three is, in the absence of reciprocity considerations, an irrelevant option. It constitutes, so to speak, a noncredible threat that should not affect behavior in a systematic way. As a consequence it should leave the distribution of earnings unaffected. Yet, our results show that this prediction of the standard model is clearly violated.

[15] Firms who rewarded workers for their excess effort chose on average $p = 0.7$. In 29.7 percent of trades with excess effort firms did not reward; i.e., they chose $p = 0$; see Table 3(a). The average reward level was 0.55.

[16] Of course, there remains the possibility of a type II error.

Table 5. *Realized gains from trade*

treatment	firms	workers	sum
2-stage	994	4398	5392
3-stage	4509	3067	7576

Result 9: In the two-stage treatment workers earn on average more than firms while the opposite holds in the three-stage treatment.

This result is best illustrated in Table 5, which shows total effective earnings in experimental money units over all twelve periods of workers and firms, respectively. In the three-stage treatment firms are considerably better off than in the two-stage treatment and better off than workers. For workers the opposite is true: They are in a better position in the two-stage treatment than in three-stage treatment and earn less than firms in the three-stage treatment. The total of realized gains from trade, however, is considerably higher in the three-stage treatment than in the two-stage treatment.

Figures 5(a) and 5(b) illustrate these facts by showing periodwise profits in the two treatments. From this we conclude that threats that are not credible in the absence of reciprocity are indeed capable of affecting the distribution of income when people are motivated by reciprocity considerations. Contrary to what the standard model predicts, having the "last word " is a valuable option in our experiments.

5 Interpretation and concluding remarks

The data of our experiments exhibit a pattern that clearly violates the predictions of the standard approach and that is in accordance with a reciprocity-based approach. Firms persistently offer contracts with large rents and non-incentive-compatible effort requirements. In the two-stage design they try to elicit workers' reciprocal responses by offering higher rents when they desire higher effort levels. Workers, in turn, behave reciprocally by choosing higher effort levels in cases in which they are offered higher rents. Workers underprovide effort relative to $\hat{e}$, but this underprovision is much smaller than predicted by the standard approach. In the three-stage design we observe that firms respond reciprocally and that workers (partly) anticipate firms' responses correctly. Moreover, firms' opportunity to punish and reward has a large effect on workers' shirking behavior. In many instances workers even overprovide effort to elicit rewards.

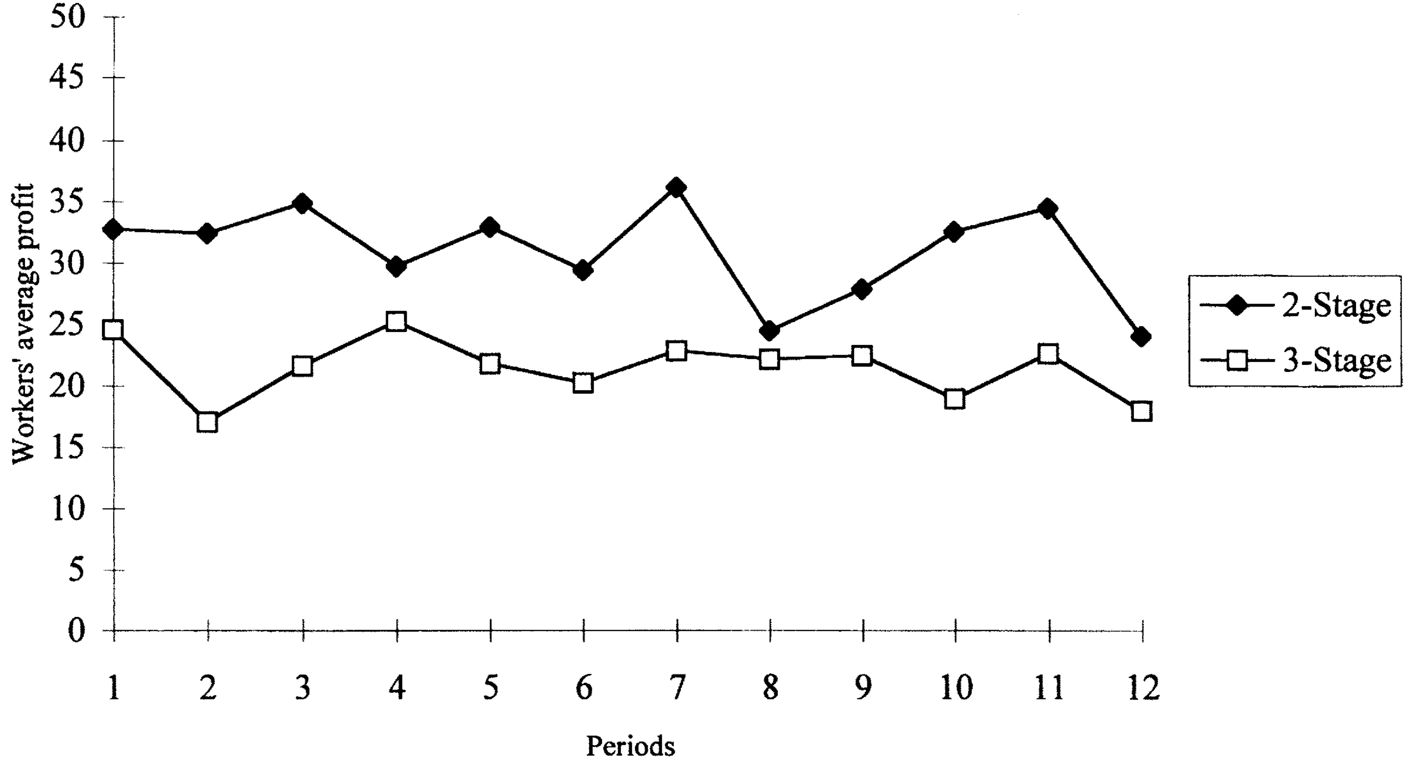

Figure 5(a). Worker's profit in the two- and three-stage treatment

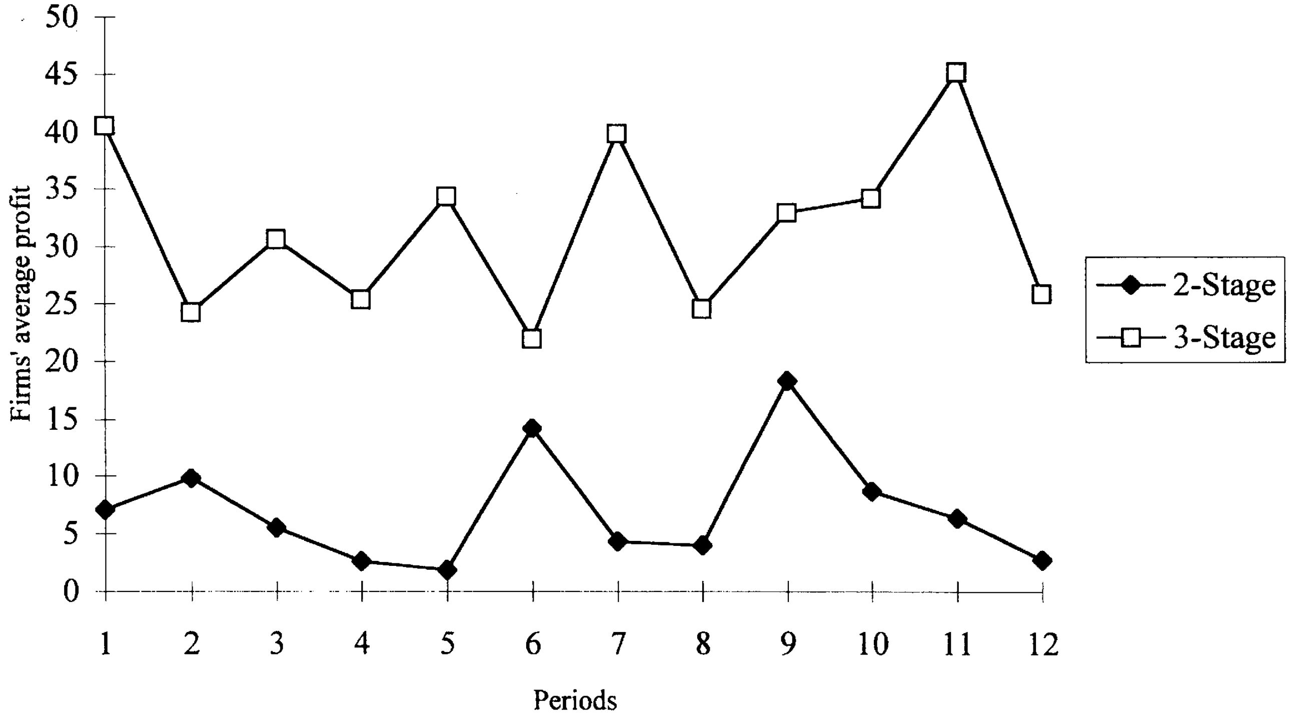

Figure 5(b). Firms' profit in the two- and three-stage treatment

These results indicate that, on average, both firms and workers try to elicit reciprocal responses. They anticipate such responses and, given the opportunity, themselves respond reciprocally. Overall these behavioral and experimental patterns result in a large increase in the total gains from trade relative to the standard prediction.

According to our interpretation the data indicate a true willingness to pay for reciprocity; that is, subjects exhibit a preference for acting reciprocally. An important objection to our interpretation concerns the fact that we had twelve market periods in each treatment. In principle, this could create opportunities for strategic and reputational spillovers across periods. However, we have taken great care in preventing such spillovers by enforcing strict anonymity between trading partners. Because of our anonymity requirements it was definitely impossible in our design that *individual* firms or workers developed a reputation. Nor was it possible that the actions of any *specific* firm or worker in previous periods could be rewarded in the present period.

A firm could, for example, not reward a high effort in period t by a high job rent offer in period $t + 1$ because it did not know the worker's identity in period t; nor could the firm address the offer in $t + 1$ to any specific worker. Nonetheless, in case that firms – for whatever reason – respond to high effort in t with high rent offers in $t + 1$ workers' effort choices could be interpreted as an investment in *group* reputation. A worker who chooses high effort levels would then provide a public good because he induces firms to make generous offers to the group of workers. This behavior is, of course, also incompatible with conventional theory because it requires nonselfish cooperation among workers. Moreover, as we will argue, the desire to establish a group reputation through nonselfish cooperation among workers cannot account for one of the major regularities of our data, that is, for reciprocal effort choices.

In our view, the fact that workers respond reciprocally to the *current* job rent is evidence against the group reputation hypothesis. Suppose, for a moment, that firms respond positively to last period's effort. Under these conditions workers should *not* respond reciprocally to the current job rent if they want to induce high future rents. If they choose low effort levels in response to low current job rents they cause low rent offers in the next period which (in the case of reciprocal effort choices) give rise to low future offers, etc. Thus, the desire to induce high future job rents by high present effort levels requires *unconditionally* high effort levels, which are, however, not observed.

An additional argument against the group reputation hypothesis arises from the evolution of effort over time. Since toward the end of the experiment the value of a group reputation decreases continuously we should observe that effort levels converge toward the minimum level of $e = 0.1$. As Figure 2, however, shows, there is no such downward trend. Even in the final period the average effort is significantly above the minimum effort level. Therefore, the data do not support the group reputation hypothesis.

Further evidence against the group reputation hypothesis comes from a series of experiments conducted by Fehr, Kirchler, Weichbold, and Gächter (FKWG) (1997). In their paper the results of a one-shot reciprocity treatment are reported. The FKWG design has the following features: Ten firms interact with ten workers over ten periods, but each firm is matched bilaterally with each worker only once. This matching procedure is common knowledge. A firm makes a wage proposal to a worker. If the worker accepts he has to choose effort and bears costs $c(e)$ according to Table 1. If he rejects the offer, both players earn zero. Because of the one-shot nature of these experiments it never pays for a subject to invest anything in group reputation. FKWG also conducted competitive market experiments with reciprocation opportunities (like our two-stage treatment) to allow for a comparison of the bilateral one-shot experiments with the competitive market experiments. Their results indicate that workers also respond reciprocally in a one-shot situation. Workers' response pattern in the one-shot situation is rather similar to their behavior in a competitive market with reciprocation opportunities. This indicates that group reputation is – if at all – not very important in the competitive market with reciprocation opportunities.

In our view the regularities of our data indicate that the neglect of reciprocal behavior may induce economists to make wrong predictions. Moreover, they may give wrong advice regarding the design of contracts and institutions. It may well be that, instead of aiming at stronger pecuniary incentives, or improving the enforcement technology, increasing the scope for reciprocal interactions is a better or an equally good means to improve the performance of agents. Likewise, it may turn out that traditional pecuniary incentives are in conflict with reciprocal incentives. Strengthening of pecuniary incentives may weaken reciprocal incentives and vice versa. An example for this is provided by Fehr, Gächter, and Kirchsteiger (1997). They report the results of experiments that are similar to the ones discussed in this chapter. The experiments differ, however, in one important respect. In Fehr, Gächter, and Kirchsteiger (1997) firms could credibly threaten ex ante to fine shirking workers. That is, their contract offers could condition wage reductions on the verifiable underprovision of effort. It turns out that in the two-stage design the opportunity of explicitly threatening a fine leads to *lower* effort levels. Thus, it seems that the introduction of traditional pecuniary incentives into a reciprocal relationship may weaken *total* incentives for effort provision. Of course, more evidence is needed. But the strong patterns of reciprocal behavior in our data suggest that the interaction between pecuniary incentives and incentives for reciprocal behavior deserves further empirical and theoretical study. A very good example for theoretical progress in this area is Rabin's (1993) model of fairness.

Appendix: A summary of our experimental procedure

Worker and firms in two different rooms; 12 periods plus a trial period

1st stage:

1. Firms choose wage $w \in [0, 100]$ and desired effort $\hat{e} \in [0.1, 1]$.
2. Workers accept in randomly determined order.

2nd stage:

3. Workers privately determine actual effort $e \in [0.1, 1]$.

3′. *(In three-stage design only)* Worker privately states expected punishment/reward level.

4. Firm is privately informed about the worker's actual effort level.

Payoffs are calculated in the two-stage treatment.

3rd stage:

5. Firm privately chooses punishment/reward level $p \in [-1, 1]$.
6. Worker is privately informed about the punishment/reward level.

Payoffs are calculated in the three-stage treatment.

REFERENCES

Adams, J. St. (1963). "Wage Inequities, Productivity and Work Quality," *Industrial Relations* 3: 9–16.

———(1965). "Inequity in Social Exchange," in Leonhard Berkowitz, ed., *Advances in Experimental Psychology 2*, pp. 267–99. New York: Academic Press.

Agell, J., and P. Lundberg (1995). "Theories of Pay and Unemployment: Survey Evidence from Swedish Manufacturing Firms," *Scandinavian Journal of Economics* 97: 295–308.

Berg, J., J. Dickhaut, and K. McCabe (1995). "Trust, Reciprocity and Social History." Discussion paper, University of Minnesota.

Bewley, T. (1995). "A Depressed Labor Market as Explained by Participants," *American Economic Review* (Papers and Proceedings) 85: 250–54.

Blinder, A., and D. Choi (1990). "A Shred of Evidence on Theories of Wage Stickiness," *Quarterly Journal of Economics* 105: 1003–16.

Camerer, C., and R. Thaler (1995). "Ultimatums, Dictators, and Manners," *Journal of Economic Perspectives* 9: 209–19.

Cialdini, R. (1993). *Influence: The Psychology of Persuasion.* New York: William Morrow.

Fehr, E., S. Gächter, and G. Kirchsteiger (1997). "Reciprocity as a Contract Enforcement Device: Experimental Evidence," *Econometrica* 65(4): 833–60.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993). "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics* 108(2): 437–60.

Fehr, E., G. Kirchsteiger, and A. Riedl (1997, forthcoming). "Gift Exchange and Reciprocity in Competitive Experimental Markets," *European Economic Review.*

Fehr, E., E. Kirchler, A. Weichbold, and S. Gächter (1997, forthcoming). "When Social Norms Overpower Competition – Gift Exchange in Experimental Labour Markets," *Journal of Labor Economics.*

Güth, Werner, and Reinhard Tietz (1990). "Ultimatum Bargaining Behavior – a Survey and Comparison of Experimental Results," *Journal of Economic Psychology* 11: 417–49.

Kahneman, D., J. Knetsch, and R. Thaler (1986). "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review* 76(4): 728–41.

Levine, D. (1993). "Fairness, Markets, and Ability to Pay: Evidence from Compensation Executives," *American Economic Review* 83: 1241–59.

Rabin, M. (1993). "Incorporating Fairness into Game Theory and Economics," *American Economic Review* 83: 1281–1302.

Roth, A. E. (1995). "Bargaining Experiments," in A. E. Roth and J. H. Kagel, eds., *Handbook of Experimental Economics.* Princeton, N.J.: Princeton University Press.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir (1991). "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review* 81: 1068–95.

Siegel, S., and N. J. Castellan, Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill.