THE HIDDEN COSTS AND RETURNS OF INCENTIVES—TRUST AND TRUSTWORTHINESS AMONG CEOs

Ernst Fehr
University of Zurich and Collegium Helveticum

John A. List
University of Maryland and NBER

Abstract
We examine experimentally how Chief Executive Officers (CEOs) respond to incentives and how they provide incentives in situations requiring trust and trustworthiness. As a control we compare the behavior of CEOs with the behavior of students. We find that CEOs are considerably more trusting and exhibit more trustworthiness than students—thus reaching substantially higher efficiency levels than students. Moreover, we find that, for CEOs as well as for students, incentives based on explicit threats to penalize shirking backfire by inducing less trustworthy behavior—giving rise to hidden costs of incentives. However, the availability of penalizing incentives also creates hidden returns: if a principal expresses trust by voluntarily refraining from implementing the punishment threat, the agent exhibits significantly more trustworthiness than if the punishment threat is not available. Thus trust seems to reinforce trustworthy behavior. Overall, trustworthiness is highest if the threat to punish is available but not used, while it is lowest if the threat to punish is used. Paradoxically, however, most CEOs and students use the punishment threat, although CEOs use it significantly less. (JEL: C91, C92, J30, J41)

1. Introduction

Trust and trustworthiness are ubiquitous in human life. Most economic transactions require trust and trustworthiness because it is rarely the case that all dimensions of a transaction can be contractually specified and enforced. Arrow (1972, p. 357) neatly expressed this idea three decades ago: "Virtually every commercial transaction has within itself an element of trust.... It can be plausibly argued that much of the economic backwardness in the world can be explained by a lack of mutual confidence."

This paper examines experimentally how people provide and respond to incentives in situations requiring trust and trustworthiness. Our data set is unique

Acknowledgments: Ernst Fehr acknowledges support from the Swiss National Science Foundation under project number 12-67751.02 and from the MacArthur Foundation Network on Economic Environments and the Evolution of Individual Preferences and Social Norms. List acknowledges support from the University of Arizona. Jonathan Alevy provided helpful comments.

E-mail addresses: Fehr: efehr@iew.unizh.ch; List: Jlist@arec.umd.edu

Journal of the European Economic Association September 2004 2(5):743–771

in that our subject pool is comprised of Chief Executive Officers (CEOs) as well as students. This provided us with the rare opportunity to examine trust and trustworthiness among CEOs under controlled conditions.[1] Economic experiments are often criticized because they are typically based on observing the behavior of undergraduate students. This can be problematic because students' behavior may not be representative of behavior in naturally occurring environments, where selection effects may have created distinct populations of economic decision makers—e.g., CEOs may be characterized by particularly selfish preferences. This criticism may be quite relevant in the domain of trusting and trustworthy behavior because trustworthiness often requires behavior that is at odds with one's own material self-interest. In this case we would expect to observe less trust and trustworthiness among CEOs in situations where trustworthiness requires nonselfish behavior. However, in sharp contrast to this conjecture, in our experiments CEOs exhibited much more trusting and trustworthy behavior than students and, as a consequence, they achieved substantially higher efficiency levels in their transactions.

Moreover, our results indicate that among CEOs, as well as among students, there are hidden costs and hidden returns of incentives. These costs and returns are hidden in the sense that they escape our attention if our reasoning is based on the assumption that people are exclusively self-interested. We show that the *use* of the explicit threat to sanction shirking backfires by inducing less trustworthy behavior; accordingly, incentives that explicitly threaten to penalize shirking may involve hidden costs. In recent years several economists have focused attention on similar phenomena (e.g., Frey and Oberholzer-Gee 1997; Kreps 1997; Benabou and Tirole 2002). However, our results also indicate that the *availability* of the sanctioning threat can be quite productive—giving rise to hidden returns of incentives. If principals *voluntarily* refrain from using the punishment threat when it is available, agents exhibit significantly more trustworthiness than if the punishment threat is not available. Thus, if agents face no punishment threat, the mere fact that the principal could have used the punishment option affects the agent's trustworthiness in a positive manner. This finding suggests that the deliberate nonuse of the threat is perceived as a particularly trusting act that is reciprocated with a particularly trustworthy act. Exhibiting trust, to some degree, seems to generate trustworthiness that rationalizes the initially exhibited trust. In other words, trust and trustworthiness reinforce each other.

1. Experimentalists recently have made a concerted effort to examine the behavior of higher-level decision makers. For example, Cooper et al. (1999) examined the ratchet effect with middle and upper level Chinese managers; Hannan, Kagel, and Moser (2002) studied gift exchange among MBA students; Camerer, Ho, and Chong (2003, 2004) reported data from a CEO subsample in a beauty contest game; and Haigh and List (2004) studied investment decisions made by traders from the Chicago Board of Trade. The interested reader should see Ball and Cech (1996) for a discussion of subject pool effects.

In view of this result one might expect that rational principals do in fact refrain from using the punishment option. It turns out, however, that only 40% of the CEOs and 20% of the students do not use the threat when it is available. Since using the threat reduces principals' earnings substantially, the frequent use of the punishment threat constitutes a puzzle that warrants further investigation.

Our preferred interpretation of the hidden costs and returns of incentives is in terms of reciprocity. Reciprocity means that people respond to acts that are perceived as kind in a kind manner and to acts that are perceived as hostile in a hostile manner (Rabin 1993; Dufwenberg and Kirchsteiger 2004; Falk and Fischbacher 1999). It seems plausible that *explicit* threats to sanction shirking may be perceived as hostile per se and convey an element of distrust. Reciprocal agents will respond to this hostility with increased shirking. Moreover, refraining from the use of a saliently available explicit threat may be perceived as a particularly kind and trusting act, which could explain why the deliberate nonuse of the threat led to less shirking compared to the exogenous absence of the incentive. Our interpretation implies that not all kinds of incentives involve hidden costs or returns because not all of the incentives will be perceived as hostile. Yet, our interpretation also means that the psychological message that is conveyed by incentives—whether they are perceived as kind or as hostile—has important behavioral effects.[2]

We believe that our results are relevant beyond the context of our experiments. This is so because in our context social preferences such as altruism, fairness, or reciprocity are likely to be a major source of trusting and trustworthy behavior, and these motives have been shown to be relevant in many other strategic situations as well (for overviews see, e.g., Camerer 2003; Fehr and Schmidt 2003; Sobel 2002). Therefore, to the extent that our results arise from interactions between social preferences and economic incentives, there is reason to believe that these interactions are also relevant in other strategic games.

There is also field evidence indicating that our results may be relevant in the workplace. In his book on wage rigidity, Bewley (1999, p. 431) reports that managers are well aware of the fact that workers have many opportunities to take advantage of employers, which renders worker morale important. Good morale means that workers are trustworthy, i.e., they also perform well in situations where they could shirk without being detected. To achieve and sustain a trustworthy workforce, punishment "should seldom be used to obtain cooperation." Therefore, "good management practice uses punishment and dismissal

2. Evidence in Fehr and Gächter (2002) supports this view. They show that the positive and negative framing of an incentive has a strong impact on effort in situations requiring trustworthiness. If the incentive is framed as a pay cut in case of shirking, agents' effort is much lower compared to an economically identical incentive that is framed as a bonus that is paid in addition to a (lower) base wage when the required performance threshold is met.

largely to deter and weed out bad characters and incompetents and to protect the company from malefactors."

The same forces that explain our data patterns may also explain why so few marriages are accompanied by prenuptial agreements. It seems plausible that prenuptial agreements are likely to introduce distrust into a marriage because they require detailed discussions and specifications of what will happen in the case of a divorce. As a consequence they may do more harm than good. Moreover, it also seems likely that being trusted is, in itself, valuable for the trustee. Including contingencies about what will happen if one party fails to abide by the contract is likely to be taken as an indication of distrust and perhaps even hostility, which in turn may trigger what the prenuptial agreement attempted to avoid—a lack of mutual trust and cooperation.[3] If it is true that explicit haggling and bargaining about the concrete terms of a marriage sows the seeds for malfunctioning, then the existence of marriage and divorce laws can be interpreted as a remedy for a market failure because these laws free the parties from the necessity to bargain ex ante about the terms of their marriage.

In view of the potential relevance of our results for a wide variety of contexts, we believe that by taking into account the effects we have documented, the economic theory of contracts and incentives will progress further. This theory has progressed substantially during the last two decades, but there are still crucial gaps in our knowledge about the effects of contracts and incentives (for excellent reviews of the empirical evidence, see Prendergast 1999 and Chiappori and Salanie 2003).

Recently, Gneezy and Rustichini (2000a, 2000b) and Bohnet, Frey, and Huck (2001) reported interesting evidence indicating counterproductive effects of incentives. Our study differs from these papers along several important dimensions. To our knowledge, our study is the first one documenting the hidden returns arising from the availability of incentives. We also believe that our study is unique in its examination of subject pool effects—we do not know of any other study that examines how CEOs use and respond to incentives under controlled conditions, and whether CEOs exhibit more or less trust and trustworthiness than the typical experimental subject pool—undergraduate students. Moreover, the mentioned authors do not study the relational aspect of incentives, that is, the impact of the incentive that arises from the fact that the principal can threaten to punish or can forgo threatening to punish the agent when given the opportunity.[4] In contrast,

3. Business relations also may not be immune to such effects. In a classic paper, Macaulay (1963) reports that "detailed negotiated contracts can get in the way of creating good exchange relationships between business units." Likewise, Sitkin and Roth (1993, p. 376) assert that "legalistic remedies can erode the interpersonal foundations of a relationship they are intended to bolster because they replace reliance on an individual's good will with objective, formal requirements."

4. In Bohnet, Frey, and Huck (2001) the principal chooses whether to enter a contract or not to enter a contract. Then the agent has the choice to honor the contract or not to honor the contract—the

our study focuses on this aspect by implementing experimental treatments that vary the principal's capability of imposing a punishment.[5]

The remainder of this study is structured as follows. In Section 1 we describe our experimental design, our subject pool, and the details of the experimental parameters and procedures. In Section 2 we report our results. Section 3 summarizes and concludes the paper.

2. Experimental Design and Procedures

To study the hidden costs and returns of incentives in situations requiring trust and trustworthiness we used two versions of a Trust game.[6] In both versions a principal (Actor 1) and an agent (Actor 2) were paired anonymously and the identity of the transaction partner was never revealed to the subjects.[7] Both the principal and the agent received an endowment of ten experimental money units called "shanks." At the end of the experiment the amount of shanks earned was exchanged into real money (U.S.$) according to a publicly known exchange rate (this represented their payoff as there was no show-up fee). In the *Trust* treatment, the principal could transfer $x \in \{0, 1, 2, \ldots, 10\}$ shanks to the agent. For every shank transferred, the agent received three shanks. In addition, the principal had to announce a "desired back-transfer" $\hat{y} \in \{0, 1, 2, \ldots, 3x\}$ from the agent. When the principal had made his decision $(x, \hat{y})$, the agent was informed about this two-tuple and then chose the actual level of the back-transfer $y \in \{0, 1, 2, \ldots, 3x\}$. The principal's payoff in the Trust treatment was given by $\Pi^p = 10 - x + y$, whereas the agent's payoff was defined as $\Pi^a = 10 + 3x - y$. Thus, the desired back-transfer $\hat{y}$ was not payoff-relevant.

---

experimenter exogenously determines the punishment technology. In Gneezy and Rustichini (2000a, 2000b) the agents also face incentives that are determined by the experimenters.

5. Our study also differs substantially from the experiments conducted by social psychologists on the undermining of intrinsic motivation by extrinsic rewards. This literature started with Deci (1971), and has led to several metastudies (see, e.g., Cameron and Pierce 1994; Deci, Koestner, and Ryan 1999). Trust and trustworthiness plays no role in these experiments, whereas it is key in our context. Moreover, in these psychology experiments subjects faced exogenous incentives (in the form of monetary rewards) set by the experimenter.

6. There is now a large literature on so-called Trust and Gift Exchange games starting with Fehr, Kirchsteiger, and Riedl (1993), and Berg, Dickhaut, and McCabe (1995)—see Camerer (2003), Fehr and Schmidt (2003) and Sobel (2002) for reviews. In general, these games are sequential social dilemma games in which trusting or trustworthy behavior enhances the total payoff of the parties involved, but in which individuals face monetary incentives inhibiting trust and trustworthiness. Our general design is similar to Fehr and Rockenbach (2003), who study the behavior of students in a Mensa experiment.

7. Our experimental procedures implemented a single-blind design that ensures that subjects do not know each other's identities and that the decisions made by a pair of actors are known only to the pair and the experimenter. At the end of the experiment, subjects are privately paid what they earned during the experiment.

In the other treatment, however, which we denote as the *Trust with Punishment* (TWP) treatment, the desired back-transfer was important. In TWP, the principal also had to make a decision about $x \in \{0, 1, 2, \ldots, 10\}$ and $\hat{y} \in \{0, 1, 2, \ldots, 3x\}$. In addition to $(x, \hat{y})$, he decided whether to impose a fixed fine of $f = 4$ that had to be paid by the agent if $y < \hat{y}$. Thus, in TWP the principal could threaten to punish the agent in case of malfeasance. In the instructions we avoided value-laden terms; that is, we did not use the word "punishment" or "fine." Instead, we spoke of a "conditional payoff cut."[8] After the principal had made his decision, which was constrained to be either $(x, \hat{y}, f = 0)$ or $(x, \hat{y}, f = 4)$, the agent was informed about the decision and had to choose y. The principal's payoff in TWP was given by $\Pi^p = 10 - x + y$ irrespective of whether the agent was fined or not. The fine was not given to the principal. The agent's payoff was given by $\Pi^a = 10 + 3x - y - 4$ if the principal imposed the fine and the agent chose $y < \hat{y}$. If the principal did not impose the fine or if the agent chose $y \geq \hat{y}$, the agent did not have to pay a fine; he earned $\Pi^a = 10 + 3x - y$.

If both the principal and agent are selfish, and if the principal anticipates the agent's selfishness, the predictions for our two treatments are straightforward. In the Trust treatment the selfish agent chooses $y = 0$ irrespective of the transfer x. Hence, a rational and selfish principal chooses $x = 0$. There is no precise prediction about the desired back-transfer $\hat{y}$ because $\hat{y}$ does not affect the payoffs. In the TWP treatment there are subgame perfect equilibria in which the principal can induce a selfish agent to pay back $y = 3$ or $y = 4$ by imposing a fine. For instance, if the principal transfers $x = 1$, demands $\hat{y} = 3$ and imposes $f = 4$, the agent's best response is $y = 3$. In this case the principal earns $\Pi^p = 10 - 1 + 3 = 12$ whereas the agent earns $\Pi^a = 10 + 3 - 3 = 10$. Likewise, if the principal transfers $x = 2$, demands $\hat{y} = 4$, and imposes $f = 4$, $y = 4$ is a best response. The principal then earns $\Pi^p = 10 - 2 + 4 = 12$ and the agent also earns $\Pi^a = 10 + 6 - 4 = 12$. It is obvious that in TWP the principal can never enforce more than $y = 4$ by imposing the fine. Yet, by choosing $f = 4$ his enforcement power exceeds the case of $f = 0$. Thus, in TWP the principal will always use the fine, and this should yield a higher payoff than under the Trust treatment, and compared to $f = 0$ in the TWP treatment. In addition, if the principal chooses $f = 0$ in the TWP treatment, the agent should respond in the same manner as in the Trust treatment where fines were excluded by design.

The only difference between the Trust and the TWP condition is the availability of the fine. The back-transfers of the agents in the Trust treatment provide us with a baseline measure of trustworthiness. Positive back-transfers in the Trust treatment may be due to the agents' preferences for inequity aversion, reciprocity

8. Appendices A, B, C, and D contain copies of the instructions.

or altruism.[9] By comparing these back-transfers with the back-transfers in TWP, we can study the effect of the availability and the actual use of the incentive on back-transfers. Likewise, by comparing principals' transfers across Trust and TWP conditions, we can examine the effect of incentive use and incentive availability on the trusting behavior of the principals. The comparison between the Trust condition and those interactions in the TWP condition in which the principal chooses $f = 0$ is particularly interesting because in economic terms the situation faced by the agents is identical in these cases. However, from a psychological viewpoint it may make a difference whether the absence of a threat is exogenously determined or endogenously chosen by the principals because the deliberate nonuse of a saliently available threat may be perceived by the agents as a particularly trusting act. Thus, by comparing back-transfers in the trust condition with back-transfers in the TWP interactions with $f = 0$, we can study whether trust breeds trustworthiness.

Since one important purpose of our project was to compare the principals' use of the incentive and the agents' response to the incentive across CEOs and students, both subject pools participated in the Trust and the TWP treatment. In the student treatments, one shank was worth \$0.20, and in the CEO treatments one shank was worth \$2.00. The different exchange rates served the purpose to control for stake effects across subject pools. We hypothesized that because CEOs have a higher income than students they need to face higher stake levels to take the experiment seriously. Postexperiment interviews revealed that both students and CEOs took the experiment very seriously.[10]

We recruited 126 subjects for our student treatments from the undergraduate student body at the University of Costa Rica. Each treatment was run in a large classroom on the campus of the University of Costa Rica. To ensure that decisions remained anonymous the subjects were seated far apart from each other. The CEO subject pool included 76 CEOs from the coffee beneficio (coffee mill) sector who were gathered at The Costa Rica Coffee Institute's (ICAFE) annual conference in March 2001.[11] The conference is funded by ICAFE and presents

9. There are now many papers that model these kinds of social preferences. On reciprocity see, e.g., Rabin (1993) or Falk and Fischbacher (1999), on inequity aversion see Fehr and Schmidt (1999) or Bolton and Ockenfels (2000), and on altruism see Andreoni and Miller (2002).

10. In an experimental session, CEOs earned on average \$65; students' average earnings were approximately \$5.65. A session lasted roughly 45–60 minutes. To put these earnings figures into perspective, note that our students could have earned about \$2 per hour in a good alternative job, and a large cup of coffee costs about \$0.25 on campus.

11. ICAFE was created in 1948, and is a semi-autonomous institution in charge of providing technical assistance, undertaking field research, supervising receipts and processing of coffee, and recording export contracts. Note that some of the data were gathered after the conference because of extenuating circumstances (there was a march/rally the day of the conference and several CEOs left the conference early or did not show up). Data gathered in this manner are not significantly different from data gathered at the conference, so we pool the data. Also, it is important to stress that we did not collect complete demographic data on CEOs. Nevertheless, it is clear that the CEOs

the CEOs with information related to the most recent technological advances in the coffee processing sector, regulations within Costa Rica as well as abroad, and general market conditions, among other agenda items. Each of the CEO treatments was run in a large room on-site at the institute. As in the case of the students, communication between the subjects was prohibited and the CEOs were seated such that no subject could observe another individual's decision.

At the beginning of the experiment, subjects signed a consent form in which they acknowledged their voluntary participation in the experiment and agreed to abide by the rules of the experiment. There were two types of sessions. In a Trust–TWP session, subjects had to participate first in the Trust treatment and then in the TWP treatment. To control for possible spillover effects across treatments within a session, we reversed the treatment ordering in the TWP–Trust sessions.[12] Each subject participated in only one type of session, and within a session each subject had the same role in the Trust and the TWP treatment. In total we conducted two sessions with the CEOs (one Trust–TWP and one TWP–Trust session) and four sessions with the students (two Trust–TWP and two TWP–Trust sessions).

A few noteworthy aspects of our experimental design merit further consideration. First, principals made their decisions on a decision sheet. When the principals had made their decisions the experimenters collected the decision sheets and gave them to (different) randomly selected agents. Then the agents made their decision on the decision sheet. Once all agents had made their decision the experimenters again collected all decision sheets and informed the principals about the decision of their agent. Second, the experimental instructions were first written in English and then translated into Spanish. This translation was performed by a Costa Rican expert. To control for translation biases, a different translator located in Arizona then translated the Spanish instructions back into English (see the *Journal's* web site for the Spanish instructions).

3. Experimental Results

This section presents our experimental results. Since we find that ordering effects are not important we pool the data in the following empirical analyses.[13] Our first result relates to the comparison between the behavior of CEOs and students.

were overwhelmingly more male, older, wealthier, etc. If our goal is to establish a CEO-student difference, it is necessary to create a parallel sample of non-CEOs matched for comparability. This is beyond the scope of our study, as we are mainly interested in testing for behavioral differences across two very different groups.

12. During the first treatment the subjects did not know that there would be a second treatment in the session. After the first treatment was completed, subjects were informed that another experiment would take place. They were also informed that the second experiment was the final one, and that they would be matched with a new partner in this experiment.

13. In addition, we explicitly control for ordering effects in our regressions below (see footnote 17).

A major criticism levied against experimental results concerns the fact that most economic experiments are conducted with students. This may be problematic for two reasons. First, students may not be a representative subject pool for the overall population. Second, due to selection effects, particular people, who do not behave like students, may be overrepresented in certain sectors of the economy. Both reasons may constitute obstacles for generalizing the results gained from student experiments to other environments. Our first result addresses this issue.

RESULT 1. *CEO principals transfer more money than students. Moreover, for any given transfer level, CEO agents pay back more money than students.*

Result 1 can be taken as an indication that CEOs are more trusting and exhibit more trustworthiness than students. They display a higher degree of trusting behavior because they make themselves more vulnerable to exploitation by transferring more money, and they exhibit more trustworthiness because, for identical transfer levels, they send back more money than students. Support for Result 1 can be found in Tables 1 and 2 and Figures 1 and 2 (all numbers in tables and figures are in shanks). Table 1 presents the average behavior of students and CEOs in both treatment types. In Table 1, boldfaced numbers indicate the average results for the CEOs, whereas the student results are in plain font. The first line in Table 1 shows that, both in the Trust as well as in the TWP condition, CEOs transfer, on average, more money to the agents than do students. This difference in transfers (investments) is significant at the $p < 0.05$ level in both treatments

TABLE 1. Comparison of trust and trust with punishment treatment

	Average (over all observations)	
	Trust	Trust with punishment
Transfer (investment) x to the agent	4.0 (2.6) **5.9 (2.3)**	5.0 (2.9) **7.3 (2.3)**
Desired payback in percent of tripled investment $\hat{y}/3x$	76.9 (41.0) **65.1 (20.5)**	69.9 (24.0) **66.1 (23.2)**
Actual payback in percent of tripled investment $y/3x$	31.6 (26.3) **44.1 (22.3)**	38.7 (33.6) **44.0 (23.3)**
Principals' payoff	10.5 (3.0) **11.8 (3.7)**	11.0 (4.9) **12.6 (4.9)**
Agents' payoff	17.5 (5.9) **20.1 (5.6)**	17.4 (7.5) **20.5 (5.1)**
Number of observations (pairs)	126 (63 pairs) **76 (38 pairs)**	126 (63 pairs) **76 (38 pairs)**

Notes: CEO data in bold. Standard deviations in parentheses. Figures are in shanks.

TABLE 2. Regression estimates for the agents' payback

	Model type					
	Students			CEOs		
Variable	OLS	Tobit	Tobit RE	OLS	Tobit	Tobit RE
Investment	0.79**	1.13**	0.94**	1.2**	1.2**	0.93**
	(5.7)	(5.8)	(5.5)	(5.7)	(5.8)	(3.7)
TWP	−0.27	−0.45	−0.26	−1.8*	−1.8*	−1.5*
	(0.36)	(0.44)	(0.33)	(1.8)	(1.8)	(1.8)
TWPN	4.6**	4.7**	3.6**	6.1**	6.2**	6.3**
	(3.5)	(2.8)	(2.4)	(4.7)	(4.8)	(4.7)
Constant	1.5*	−1.1	−0.84	1.1	1.1	2.3
	(1.7)	(0.87)	(0.71)	(0.78)	(0.80)	(1.2)
Person random effects	NO	NO	YES	NO	NO	YES
R^2	0.37	—	—	0.55	—	—
n	126	126	126	76	76	76

Notes: Dependent variable is the agent's transfer back to the principal (in shanks). t-ratios (in absolute value) are beneath coefficient estimates. Tobit RE parameter estimates are marginal effects computed at the sample means.
** Significant at the 0.05 level.
* Significant at the 0.10 level.

according to a two-sided Mann–Whitney test ($z = 3.76$ in the Trust and $z = 3.67$ in the TWP condition).[14]

The second line in Table 1 shows that in both conditions the CEOs' desired back-transfer is roughly 66% of the tripled investment. This means that on average CEO-principals proposed to divide the total amount of money that was made available through their transfers such that the payoff to the principal and the agent became equal. Table 1 also shows that the students' desired back-transfer is slightly higher in both treatments. This difference is, however, not significant at conventional significance levels according to a Mann–Whitney test. The third line in Table 1 indicates that in both conditions the agents' payback, measured as a percentage of the tripled investment, is higher for CEOs. Note that by measuring the payback as a percentage of the tripled investment we are explicitly controlling for the investment level. While the CEOs pay back roughly 44% of the tripled investment in both conditions, the students' payback varies between 32% and 39%.[15] Further support for differences in trustworthiness can be gained from Figures 1 and 2 (Figures 1b, 1c, 2b, and 2c complement Figures 1a and 2a by plotting the raw data from the treatments). The figures compare the payback of the students with the payback of the CEOs at various investment levels. Figure 1a,

14. All test results reported in the paper use two-sided alternatives.

15. In cases where principals transferred zero shanks but requested a positive return, we set the desired payback in percent of tripled investment equal to 1. In these cases where the agent sent back zero shanks, we set the actual payback in percent of tripled investment equal to 0.

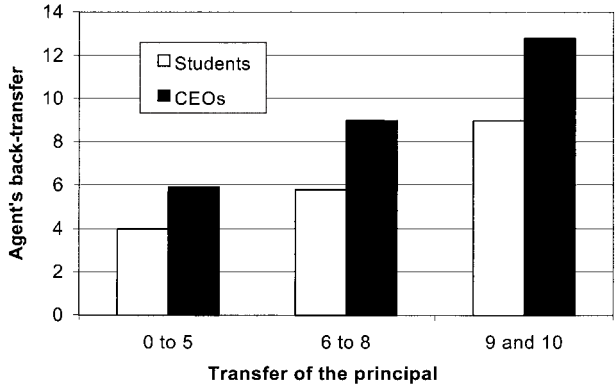

FIGURE 1a. Back-transfers of CEOs and students in the trust treatment.

which is based on data gathered in the Trust condition, shows that CEOs pay back considerably more than students at each investment interval.[16] Figure 2a, which is based on data from the TWP condition, also indicates that CEOs are more trustworthy at every given investment interval.

To explore these differences further, we estimate OLS, Tobit, and Tobit random effects regression models for the students and the CEOs. The dependent variable in the regressions is the payback of the agent, which is regressed on the principals' transfer to the agents and two dichotomous regressors. Formally, our regression models are given by

$$y_{it} = \beta_1 + \beta_2 x_{it} + \beta_3 * \text{TWP} + \beta_4 * \text{TWP} * \text{TWPN} + \omega_{it}. \qquad (1)$$

In equation (1), y_{it} represents Agent i's payback to the principal in Period t, and x_{it} denotes the principal's transfer (investment) to Agent i in Period t.[17] The dummy variable TWP equals 1 if the subject is in the TWP condition, 0 otherwise. The dichotomous variable TWPN takes on the value of 1 if the subject is in the TWP condition and does not use the punishment option, 0 otherwise. Thus, the coefficient β_4 on the interaction term TWP $*$ TWPN measures the marginal effect of not using the punishment option when the option is available, while the coefficient β_3 measures the effect, relative to the Trust condition, of being in treatment TWP *and* using the punishment option. Since there are typically quite substantial individual differences in data sets where social preferences play a role, and since our sequential design may have created dependencies between the

16. We partitioned the data into intervals because at some investment levels there are only a few observations (see Figures 1b, 1c, 2b, and 2c).

17. We also include a time dummy variable in ω_{it} that captures ordering effects. They are never significant so we suppress further discussion.

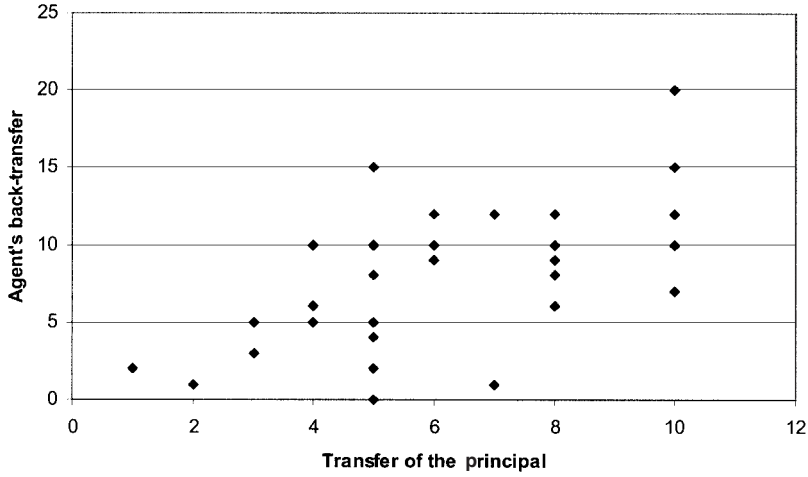

FIGURE 1b. CEO raw data summary: Trust treatment.

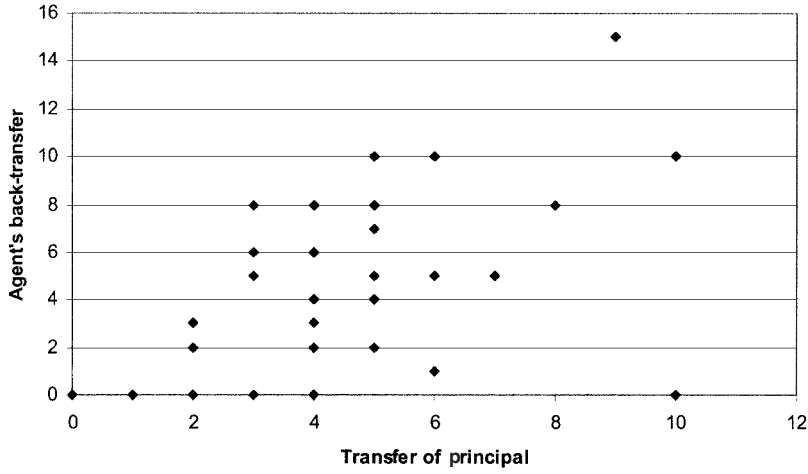

FIGURE 1c. Student raw data summary: Trust treatment.

Trust and the TWP condition, the Tobit random effects model is the appropriate specification. Nevertheless, for completeness, and to give an indication of the robustness of our results, we also report empirical results from OLS and Tobit regression models.

Table 2 contains summary regression estimates. In each regression model, empirical estimates from both the student and CEO samples indicate that the coefficient on investment is positive and significant at the $p < 0.05$ level. While

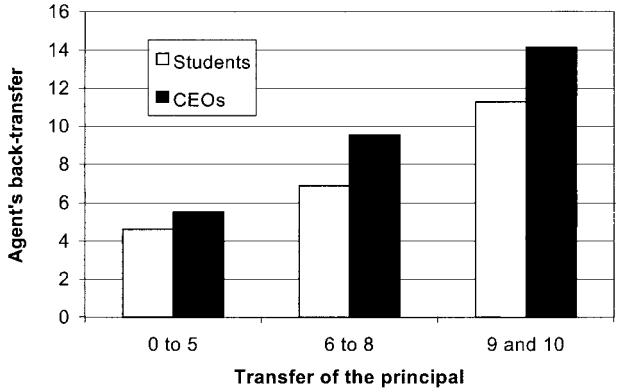

FIGURE 2a. Back-transfers of CEOs and students in the trust with punishment treatment.

in the OLS model the coefficient in the CEOs' specification is considerably larger, the null hypothesis that the investment coefficient for the CEOs is identical to the coefficient for the students in the Tobit random effects model cannot be rejected at conventional levels. This result suggests that on the margin CEOs and students behave similarly (i.e., the effect of a one-unit increase in x_{it} measured at the sampling means is isomorphic across samples).[18]

If subjects care only for their own material payoff, then the punishment option is a useful tool for principals to induce agents to transfer back some money. Thus, if the punishment option is available, the principal will always be better off in money terms if the option is used. Our next result shows, however, that in the presence of social preferences this argument may be seriously misleading.

RESULT 2. *If the punishment option is available, then agents pay back more money and principals earn more money if the option is not used.*

Result 2 suggests that the use of the punishment option generates costs that are overlooked if social preferences are neglected. Instead of increasing the amount transferred back, the punishment actually reduces the payback. A first indication in support of Result 2 is contained in Table 3, which separates the data in the TWP

18. It may also be the case that y_{it} is convex in x_{it}. We included higher-order terms in equation (1) to examine this issue and for the CEO data inclusion of these terms was not appropriate. For the student data, however, higher-order terms were statistically significant to a cubic. In these specifications, on the margin students returned less money than CEOs returned for x_{it} values greater than five (three) in the quadratic (cubic) model.

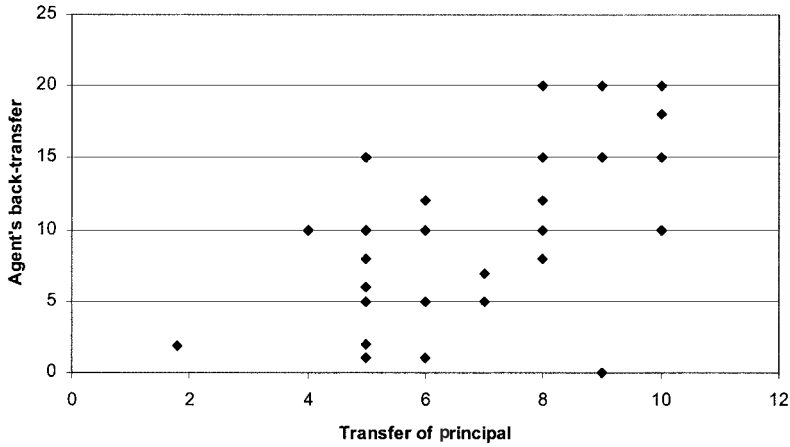

FIGURE 2b. CEO raw data summary: Trust with punishment treatment.

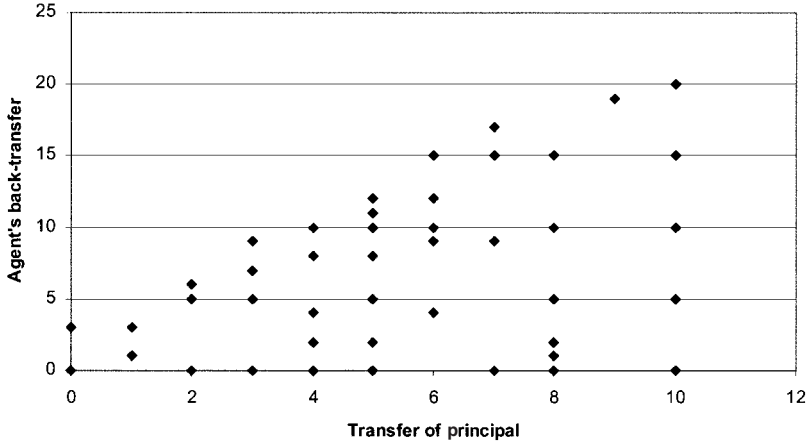

FIGURE 2c. Students raw data summary: Trust with punishment treatment.

condition according to whether the principals imposed the punishment. Boldfaced numbers in Table 3 indicate that refraining from the use of the punishment option nearly doubles the actual payback of the CEOs as a percentage of the tripled investment. If the punishment option is used, then CEOs pay back 32.7% of the tripled investment, while if the option is not used they pay back 61.4%. For students, the increase in the payback is smaller but remains quite substantial. In addition, Figures 3 and 4 show that, for any given investment interval, both CEOs

TABLE 3. Splitting up the trust with punishment treatment

	Average (over all observations)		
		Trust with punishment	
	Trust	No punishment	Punishment
Transfer (investment) x to the agent	4.0 (2.6)	7.4 (1.7)	4.4 (2.8)
	5.9 (2.3)	**8.5 (2.1)**	**6.5 (2.1)**
Desired payback in percent of tripled investment ($\hat{y}/3x$)	76.9 (41.0)	60.4 (22.3)	72.5 (24.0)
	65.1 (20.5)	**66.8 (19.3)**	**65.6 (26.0)**
Actual payback in percent of tripled investment ($y/3x$)	31.6 (26.3)	52.9 (22.1)	34.9 (35.2)
	44.1 (22.3)	**61.4 (15.6)**	**32.7 (20.5)**
Principals' payoff	10.5 (3.0)	14.1 (5.0)	10.2 (4.7)
	11.8 (3.7)	**16.5 (2.7)**	**10.0 (4.3)**
Agents' payoff	17.5 (5.9)	20.8 (6.2)	16.5 (7.6)
	20.1 (5.6)	**20.4 (4.9)**	**20.6 (5.2)**
Number of observations (pairs)	126 (63 pairs)	26 (13 pairs)	100 (50 pairs)
	76 (38 pairs)	**30 (15 pairs)**	**46 (23 pairs)**

Notes: CEO data in bold. Standard deviations in parentheses. Figures are in shanks.

and students pay back much more money if the punishment option is not utilized. To examine whether this increased payback also led to increased payoffs for the principals, we turn again to Table 3, which reveals that both for CEO principals and for student principals the material payoff is higher if the punishment option is not used.

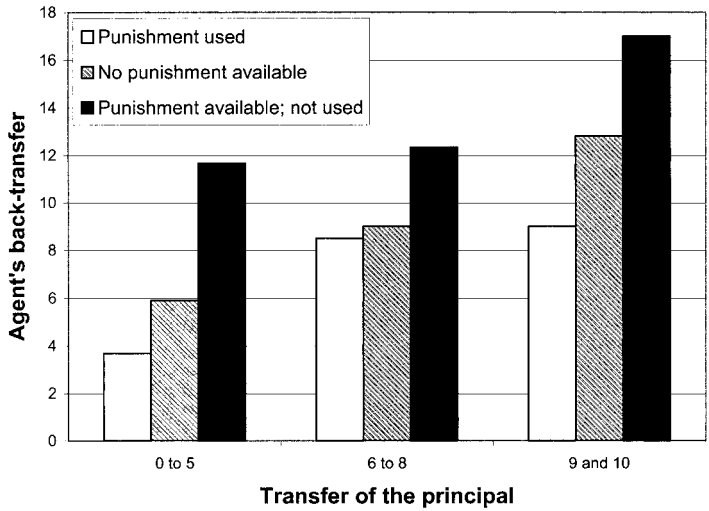

FIGURE 3. The impact of the punishment threat on CEOs' back-transfers.

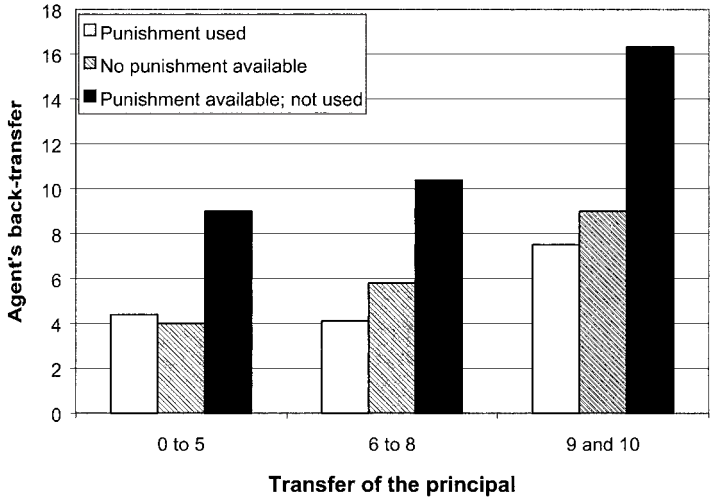

FIGURE 4. The impact of the punishment threat on students' back-transfers.

To examine whether the increase in the payback that is associated with the nonuse of the punishment option is significant, we performed several statistical tests. A Mann–Whitney test indicates that the null hypothesis of equivalent paybacks (in percent of tripled investments) can be rejected at the $p < 0.05$ level for CEOs and the $p < 0.10$ level for students. Similarly, Mann–Whitney tests of the null hypotheses that the principals' payoffs are not affected by the use of the punishment option can be rejected at the $p < 0.05$ level for both CEO and student samples. Furthermore, the regression coefficient of TWPN in Table 2 measures the increase in investment that is associated with the nonuse of the punishment option. For the CEO data, regression estimates in Table 2 indicate that not using the punishment option increases the payback significantly—approximately by six units. In the student sample the increase is roughly four units. Both of these empirical estimates are significantly different from zero at the $p < 0.05$ level.

Result 2 informs us about the hidden costs of incentives. These costs are hidden in the sense that they have generally been overlooked by standard contract theory because they are absent from models that are based on the self-interest assumption. Next we examine whether there exist hidden returns of incentives—returns that accrue from the availability of the incentive. We have already found that refraining from punishment, when the option is available, is associated with higher paybacks. This raises the question of whether the deliberate nonuse of the punishment option is better than the nonavailability of the punishment option. If it were true that one could increase one's payback by deliberately not using the

punishment option, we could speak of the hidden returns of incentives because these returns can accrue only if the incentive is available. As before, these returns are hidden in the sense that standard contract theory has neglected such effects. Our next result characterizes these hidden returns to incentives.

RESULT 3. *If in the TWP condition the available punishment option is not used, then the agents pay back more money, and the principals earn more money, than in the Trust condition where no punishment option is available.*

Support for Result 3 comes first from Table 3, which suggests that both the students as well as the CEOs pay back more money if the punishment option is deliberately not used than when it is unavailable. For instance, for the CEOs (students) the actual payback is 44.1% (31.6%) of the tripled investment in the Trust treatment, whereas in the TWP treatment the CEO agents (student agents) pay back 61.4% (52.9%) if there is no punishment threat. We find that, via a Wilcoxon signed-ranks test for matched pairs, this difference is significant at the $p < 0.05$ ($p < 0.10$) level for CEOs (students). Similar insights follow from inspection of Figures 3 and 4: not using the punishment option increases the agents' payback at every investment interval relative to not having the option available. Moreover, Table 3 also indicates that CEO principals and student principals earn more money when they deliberately do not use the punishment option, a difference that is significant at the $p < 0.05$ level for CEOs, but only marginally significant in the student sample ($p < 0.15$).

To examine whether these differences are statistically significant in our regression framework, we turn to the empirical results in Table 2. Note that the effect of the deliberate nonuse of the punishment option, relative to the Trust condition, can be measured via the summation of coefficients β_3 and β_4 in our regression model. This follows because $\beta_3 + \beta_4 * \text{TWPN} = \beta_3 + \beta_4$ if the subject is in the TWP condition and the principal did not use the punishment option (i.e., $\text{TWPN} = 1$). Table 2 reveals that for each regression model (except students Tobit RE) the sum $\beta_3 + \beta_4$ exceeds 4, suggesting that the deliberate nonuse of the punishment option, relative to the Trust condition, has a positive effect. A joint test of statistical significance suggests that the null hypothesis $\beta_3 + \beta_4 = 0$ can be rejected at the $p < 0.05$ level for both the CEO and student sample. Taken together, our results therefore suggest that there are indeed significant hidden returns of incentives: it is advantageous to have the ability of signaling that you abstain from using the punishment option compared to not having the punishment option available.

Thus far our results reveal that not using the punishment option dominates, in a material sense, both using the option and not having the option available. This raises the question of whether principals took advantage of this effect by not using the option. Our next result shows that this was not the case.

RESULT 4. *The majority of CEO and student principals use the punishment option in the TWP condition. However, CEOs use the punishment option less often.*

In total, 79.4% of the student principals (50 out of 63) chose the punishment option while only 60.5% of the CEO principals (23 of the 38) chose the punishment option. This difference is significant at the $p < 0.05$ percent level ($z = -2.05$) according to a Fisher's Exact test.

Why did the majority of student and CEO principals in the TWP condition choose the punishment option although this caused a substantial reduction in their monetary payoff? One possibility is that the principals have social preferences that induce them to choose the punishment option. For example, if principals dislike being worse off than agents, they may want to correct the payoff inequality created by a shirking agent by choosing the punishment option. Thus, if principals are sufficiently inequality averse in the sense of Fehr and Schmidt (1999) or Bolton and Ockenfels (2000), and if they anticipate a sufficiently high probability that the agent will shirk, they will use the punishment option.[19] Likewise, if principals exhibit reciprocal preferences they prefer to punish hostile actions. In our context this means that reciprocal principals prefer to punish a shirking agent even if it is costly for themselves.[20]

Another reason for the frequent use of the punishment option could be that the principals did not anticipate the negative effect on paybacks that is caused by using the option. There is some evidence (Heath 1999) indicating that people have an extrinsic incentive bias—people tend to believe that others are more motivated than themselves by extrinsic incentives. In our setting, an extrinsic incentive bias might have led the principals to favor the punishment option. Our data do not allow us to parse out the correct interpretation for the behavior of the principals. It is, however, possible to rule out rational inequality aversion as a reason for the principals' behavior. Note that if inequality-averse principals are rational they will impose the punishment only if they can rationally expect to correct the payoff inequality between principals and agents. Yet, Table 3 shows that by using the punishment option in the TWP condition, the payoff advantage of the agents *increases* from 3.9 to 10.6 for the CEOs. Thus, rational inequality-averse CEO principals have even more reason not to use the punishment option compared to

19. The Fehr-Schmidt model of inequality aversion assumes that a substantial fraction of individuals care for their own material payoff and the absolute differences between their own payoff and the payoff of relevant reference actors. In our setting the agent clearly is a relevant reference actor for the principal. The Bolton–Ockenfels model assumes that a fraction of the subjects care about the payoff ratio.

20. For models of reciprocity see, e.g., Rabin (1993); Dufwenberg and Kirchsteiger (2004); or Falk and Fischbacher (1999).

money-maximizing CEO principals. In the case of students, Table 3 indicates that the use of the punishment option changes the payoff advantage of the agents, on average, from 6.7 to 6.3. A Mann–Whitney test indicates that this small difference is not significant at conventional levels. Thus, rational inequality-averse student principals also have no reason to use the punishment option because this reduces their absolute payoff without changing the payoff differences. Therefore, only the negative reciprocity hypothesis and the extrinsic incentive bias hypothesis remain plausible candidates for explaining the principals' behavior in the TWP condition.

In view of the possibility that the majority of principals might be prone to an extrinsic incentive bias, it is interesting to ask whether the use of the punishment option in the TWP condition decreases the payback compared to the Trust condition where the incentive was not available. In addition, it is interesting to know whether principals who use the punishment option gain in material terms if they are deprived of the punishment option. Our next result addresses this question.

RESULT 5. *CEO agents pay back more money if the punishment option is not available than when it is available and used. Student agents pay back the same amount of money irrespective of whether the option is not available or available and used. Neither CEO nor student principals who use the punishment option gain in material terms if they are deprived of the option.*

Support for Result 5 comes from Tables 2 and 3. Table 3 shows that CEO agents pay back 32.7% of the tripled investment if they face a punishment threat while if this threat is not available they pay back 44.1%. Using a Wilcoxon signed-ranks test, we find that this difference is statistically significant at conventional levels ($p < 0.05$). Our regression model also confirms this result. In Table 2, the coefficient for TWP measures the effect of being in the TWP condition and using the punishment option relative to the Trust condition. The Tobit random effects specification for the CEOs suggests that the estimated coefficient is negative and significantly different from zero at the $p < 0.10$ level; a coefficient estimate of 1.5 implies that CEO agents pay back roughly 1.5 units less in the TWP if they face a punishment threat. To examine whether this led to a higher payoff for the CEO principals, we applied a Wilcoxon signed-ranks test to the principals' payoffs in row 8 of Table 3 and found that the null hypothesis of equal payoffs cannot be rejected at conventional levels ($z = 1.16$).

Table 3 indicates that student agents even decrease the payback from 34.9% of the tripled investment, if they face a punishment threat, to 31.6% percent of the tripled investment, if no threat is available. However, both a Wilcoxon signed-ranks test and the Tobit random effects regression model in Table 2 show that this payback change is not statistically significant, and that the size of the TWP-coefficient is small. Likewise the null hypothesis of equal principal

payoffs cannot be rejected according to a Wilcoxon signed-ranks test. Thus, neither the CEO nor the student principals who use the punishment option would gain in material terms if they were deprived of the possibility to use the option.

One of the outstanding features of our data is that both student and CEO agents pay back more money if the principals deliberately do not use the punishment threat (see Figures 3 and 4). In our view the notion of reciprocity provides a natural interpretation of these data. Recall that reciprocity means that agents respond in a hostile manner to actions that reveal a hostile intention. In our treatments, the threat to punish agents may reveal hostile intentions for two reasons. First, the threat per se may be perceived as hostile. Second, agents may perceive the threat as an indication of distrust. To the extent to which trusting actions are perceived as kind and distrusting actions as hostile, an explicit punishment threat may be perceived as a hostile act. Whatever the reason for interpreting someone's intention as hostile may be, if reciprocal agents perceive the explicit threat as a hostile act they are less willing to pay back money beyond the level that is dictated by pure self-interest. In this context we find one fact particularly interesting. Note that, irrespective of whether the punishment threat was used, CEO principals desired to receive a back-transfer $\hat{y}$ that, if $\hat{y}$ was met, equalized payoffs of the principal and the agent (see boldfaced figures in Line 4 of Table 3). Thus, the actions of the CEO principals implied a desire for a fair payoff distribution both when they used and when they did not use the punishment option in the TWP. Nevertheless, when the punishment option was used, the agents paid back only half as much (as a percentage of tripled investment). In our view this lends support to our interpretation in terms of the perceived hostility of the punishment threat because the different payback levels of the agents cannot be attributed to differences in the fairness of the intended payoff distribution.

The previous argument interprets the difference in agent behavior within the TWP condition in terms of reciprocity. Further evidence for the reciprocity interpretation is provided by the difference in agents' behavior between the TWP condition, when punishment is not used, and the Trust condition. In both situations principals do not use the punishment option. The only difference is that in the Trust condition they cannot use it, while in the TWP condition they voluntarily do not use it. In our view it is quite plausible that the deliberate nonuse of the punishment option is perceived as a particularly kind and trusting move. As a consequence, reciprocal agents have a particular reason to reward this move.

Of course, it would be reassuring for our interpretation if the agents did not only perceive the nonuse of the punishment threat as a particularly trusting act but if the principals in the TWP did in fact exhibit more trust when they refrained from using the threat. To examine this question we examine transfer levels in the TWP treatment and find the following.

RESULT 6. *Principals who do not punish in the TWP condition transfer more money to the agents than both principals who punish, and principals who are deprived of the punishment option.*

Table 3 shows that CEO principals in the TWP condition invested 6.5 units if they threatened to punish, while if they did not use the threat they invested 8.5 units. This difference is significant at the $p < 0.05$ level according to a Mann–Whitney test. The table also shows that student principals increased their investment from 4.4 units when they used the punishment threat to 7.4 units when they did not, which is again significant at conventional levels (Mann–Whitney test: $p < 0.05$). This indeed suggests that those principals who voluntarily refrained from the punishment threat exhibited more trust. Thus, in view of this finding, it also seems quite rational for the agents to interpret the absence of the punishment threat in the TWP condition as a particularly trusting act. It is also interesting that those principals who do not punish in the TWP condition also invest much more than the principals in the Trust condition. This effect is again significant for CEOs as well as for students ($p < 0.05$ for CEOs and for students according to a Wilcoxon signed-ranks test for matched pairs). This suggests that these principals actually exhibit more trust if they can signal their good intentions.

Finally, we consider efficiency consequences of the availability and the use of the punishment option. Efficiency is determined jointly by the actions of principals and agents. Principals' transfers determine the total pie that is available for the two parties in the Trust condition and the TWP condition if the punishment is not chosen. If, in the TWP condition, the punishment option has been chosen, paybacks of the agents determine whether the total available pie is reduced by actual punishments.

RESULT 7. *CEOs consistently achieve higher efficiency levels. Across conditions, efficiency is highest when the punishment option is available and not used and lowest if the punishment option is used. Intermediate efficiency levels prevail if the punishment threat is not available.*

We measure efficiency as a percentage of the maximum surplus that could have been generated by the two parties. In all treatment conditions the maximum surplus is 40 shanks ($80 for CEOs, $8 for students). In the Trust condition and the TWP condition without a punishment threat, this level is achieved if the principal transfers his entire endowment to the agent. If, in the TWP condition, the principal uses the punishment threat, an additional requirement for achieving the maximum surplus is that the agent pays back the desired amount of money. Support for Result 7 is provided in Table 4, which shows the efficiency levels that are reached by CEOs and students in the various conditions. The table shows that across all conditions the CEOs achieve between 5 and 12 percentage points

TABLE 4. Efficiency

	Average efficiency
Trust game	70.0% (12.9)
	79.7% (11.4)
Trust with punishment	70.9% (15.9)
	82.9% (13.4)
Trust, no punishment used	87.3% (8.3)
	92.3% (10.7)
Trust, punishment used	66.7% (14.7)
	76.6% (11.5)

Notes: Efficiency = (principal's payoff + agent's payoff)/40. CEO data in bold. Standard deviations in parentheses.

higher efficiency levels than the students. The table also shows that—irrespective of whether we examine the CEO data or the student data—efficiency is highest in the TWP condition without punishment and lowest in the TWP condition with punishment. Thus, consonant with the results highlighted previously, viewed from the efficiency perspective it is good to have the incentive available but to refrain from using it.

4. Concluding Remarks

We examined how CEOs provide and respond to incentives in situations requiring trust and trustworthiness. Our data show that CEOs exhibit considerably more trustful and trustworthy behavior than students. As a consequence, CEOs reach substantially higher efficiency levels. These results indicate that nonpecuniary motives may play a more important role in transactions among CEOs than in transactions among students. The fact that CEOs use the punishment option less often suggests that CEOs better recognize the vital role that trust plays in eliciting trustworthy behavior.

Behavior of both CEOs and students indicates that there are *hidden* costs and returns of explicit punishment threats. This result suggests that if we focus our attention exclusively on the self-interest motive when assessing the effects of incentives such effects will escape our attention. Both CEOs and students respond in a less trustworthy manner if they face an explicit punishment threat in case of shirking. However, while the use of the punishment threat causes hidden costs, the availability of the punishment threat causes hidden returns: if principals voluntarily refrain from the threat to punish, then they induce more trustworthy behavior than in a situation in which no punishment threat is available. This indicates that trustworthiness is affected by not only the absence or presence of the punishment threat but also by the reason for the absence of the threat. It seems that voluntarily refraining from the threat to punish is perceived as a particularly

trusting act, which is reciprocated with a particularly trustworthy act. Thus, in our context trust breeds trustworthiness.

Taken together it turns out that trustworthiness and efficiency is highest if the explicit punishment threat is available but not used, while it is lowest if the explicit threat is used. Despite this finding, it is not wise to make the punishment threat unavailable because the availability of the threat enables principals to signal trust by deliberately not using the threat. However, the vast majority of students and a majority of CEOs forego this opportunity of signaling trust, causing a substantial reduction in their material payoff. It is an interesting and important task for future research to examine the precise reasons behind this paradox.

Appendix A: Instructions for Trust-condition

You are actor 1

Description of your decision problem

You are a participant in the following decision-making problem. You have been randomly matched with another participant in this problem who is in another room. You will never be informed of the identity of this person, either during or after the experiment; similarly, your matched participant will never be informed about your identity. You are in the role of **actor 1** and the matched participant is in the role of actor 2. You as well as actor 2 participate only once in this decision problem. You make your decisions with the help of the decision sheet that has been handed out together with this description. Here are the rules that you and actor 2 have to obey when you make your decisions:

Endowment
At the beginning both actors receive an initial endowment of 10 shanks.

Your decision
You have to make a decision that consists of two components:

❶ **A *transfer* between 0 and 10 shanks to actor 2.**
You can transfer any amount between 0 and 10 shanks **to actor 2**. You make this decision by indicating a number between 0 and 10 in the appropriate box on your decision sheet. We will then triple this transferred amount, i.e., actor 2 receives three times the amount of shanks you transferred.

❷ **A *desired back-transfer* from actor 2.**
After you have made your transfer to actor 2 you indicate a desired back-transfer on your decision sheet. The desired back-transfer is the amount you would like to receive back from actor 2. The desired back-transfer can be any number between 0 and three times the amount you have transferred.

The decision of actor 2
Once you have fixed both components of your decision, we collect your decision sheet and give it to actor 2. In this way we inform actor 2 about your decisions. Then actor 2 can transfer any amount of the total number of shanks he received **back to you**.

Payoffs

You as actor 1 receive: 10 shanks − transfer to actor 2 + back-transfer from actor 2.

Actor 2 receives: 10 shanks + 3 × transfer from actor 1 − back-transfer to actor 1.

Exchange rate: For every shank you earn you will be paid $2 (U.S. dollars).

Appendix B: Instructions for Trust with Punishment Condition

You are actor 1

Description of a New Decision Problem

You will now participate in a new decision problem. As before you are randomly matched with another participant in another room. You are again in the role of **actor 1**. The other participant is in the role of actor 2. Notice that in this new decision problem you are matched with a **new person**, i.e., actor 2 is now a different person compared to the previous problem. Once again, you will never be informed of the identity of this person, either during or after the experiment; similarly, your matched participant will never be informed about your identity.

The new decision problem is—with one exemption—identical to the previous problem. The exemption concerns the conditional payoff cut. ***In the new problem you can impose a conditional payoff cut of 4 shanks on actor 2.*** In every other respect the problem is the same. Thus both actors receive again an initial endowment of 10 shanks.

Your decision

Again you have to indicate on your decision sheet what amount you want to transfer to actor 2 and what your desired back-transfer is. Actor 2 receives three times the amount of shanks you transferred.

In addition to the transfer and the desired back-transfer you also have to indicate on your decision sheet if you want to impose a conditional payoff cut of 4 shanks on actor 2.

- A conditional payoff cut of 4 shanks for actor 2 has the following consequences: The payoff of actor 2 will be reduced by 4 shanks if his actual back-transfer is less than your desired back-transfer. The conditional payoff cut is not due, i.e., it does not reduce the income of actor 2, if actor 2 transfers exactly your desired amount or more to you.
- If you do not impose a conditional payoff cut—the income of actor 2 will not be reduced, irrespective of how large the back-transfer of actor 2 is.

The decision of actor 2

Once you have fixed all three components of your decision, we collect your decision sheet and give it to actor 2. In this way we inform actor 2 about your decisions. Then actor 2 can transfer any amount of the total number of shanks he received **back to you**. In case that you have chosen a conditional payoff cut of 4 shanks, and if actor 2 transfers back less than what you desired, the conditional cut is due.

Payoffs

You as actor 1 receive:	10 shanks – transfer to actor 2 + back-transfer from actor 2.
Actor 2 receives:	10 shanks + 3 × transfer from actor 1 – back-transfer to actor 1 – 4 shanks (in case that a conditional payoff cut has been imposed **and** is due)

Exchange rate: For every shank you earn you will be paid $2 (U.S. dollars).

Appendix C: Instructions for Trust Condition

You are actor 2

Description of your decision problem

You are a participant in the following decision-making problem. You have been randomly matched with another participant in this problem who is in another room. You will never be informed of the identity of this person, either during or after the experiment; similarly, your matched participant will never be informed about your identity. You are in the role of **actor 2** and the matched participant is in the role of actor 1. You as well as actor 1 participate only once in this decision problem. You make your decisions with the help of a decision sheet that will be given to you after actor 1 has indicated his decision on this sheet. Here are the rules that you and actor 1 have to obey when you make your decisions:

Endowment
At the beginning both actors receive an initial endowment of 10 shanks.

The decision of actor 1
First actor 1 has to make a decision that consists of the following two components:

❶ **A *transfer* between 0 and 10 shanks to you.**
Actor 1 can transfer any amount between 0 and 10 shanks to **you**. Actor 1 makes this decision by indicating a number between 0 and 10 in the appropriate box on the decision sheet. We will then triple this transferred amount, i.e., you will receive three times the amount of shanks transferred by actor 1.

❷ **A *desired back-transfer* from you.**
After actor 1 made a transfer to you he indicated a desired back-transfer on the decision sheet. The desired back-transfer is the amount he would like to receive back from you. The desired back-transfer can be any number between 0 and three times the amount that actor 1 has transferred to you.

Your decision
Once actor 1 has fixed both components of the decision, we collect the decision sheet and give it to you. In this way we inform you about actor 1's decisions. Then you can transfer any amount of the total number of shanks you received **back to actor 1**.

Payoffs

Actor 1 receives:	10 shanks – transfer to actor 2 + back-transfer from actor 2.
You as actor 2 receive:	10 shanks + 3 × transfer from actor 1 – back-transfer to actor 1.

Exchange rate: For every shank you earn you will be paid $2 (U.S. dollars).

Appendix D: Instructions for Trust with punishment condition

You are actor 2

Description of a New Decision Problem

You will now participate in a new decision problem. As before you are randomly matched with another participant in another room. You are again in the role of **actor 2**. The other participant is in the role of actor 1. Notice that in this new decision problem you are matched with a **new person**, i.e., actor 1 is now a different person compared to the previous problem. Once again, you will never be informed of the identity of this person, either during or after the experiment; similarly, your matched participant will never be informed about your identity.

The new decision problem is—with one exemption—identical to the previous problem. The exemption concerns the conditional payoff cut. ***In the new problem actor 1 can impose a conditional payoff cut of 4 shanks on you.*** In every other respect the problem is the same. Thus both actors receive again an initial endowment of 10 shanks.

The decision of actor 1

Again actor 1 has to indicate on the decision sheet what amount he wants to transfer to you and what his desired back-transfer is. You receive three times the amount of shanks actor 1 transferred to you.

In addition to the transfer and the desired back-transfer actor 1 also has to indicate on the decision sheet if he wants to impose a conditional payoff cut of 4 shanks on you.

- A conditional payoff cut of 4 shanks has the following consequences for you: Your payoff will be reduced by 4 shanks if your actual back-transfer is less than the back-transfer desired by actor 1. The conditional payoff cut is not due, i.e., it does not reduce your income, if you transfer exactly the desired amount or more to actor 1.
- In case that actor 1 does not impose a conditional payoff cut—your income will not be reduced, irrespective of how large your back-transfer to actor 1 is.

Your decision

Once actor 1 has fixed all three components of the decision, we collect the decision sheet and give it to you. In this way we inform you about actor 1's decisions. Then you can transfer any amount of the total number of shanks you received **back to actor 1**. In case that actor 1 imposed a conditional payoff cut of 4 shanks, and if you transfer back less than actor 1's desired amount, the conditional cut is due.

Payoffs

Actor 1 receives:	10 shanks – transfer to actor 2 + back-transfer from actor 2.
You as Actor 2 receive:	10 shanks + 3 × transfer from actor 1 – back-transfer to actor 1 – 4 shanks (in case that a conditional payoff cut has been imposed **and** is due)

Exchange rate: For every shank you earn you will be paid $2 (U.S. dollars).

References

Andreoni, James and John Miller (2002). "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, 70, 737–753.

Arrow, Kenneth (1972). "Gifts and Exchanges." *Philosophy and Public Affairs*, 1, 343–362.

Ball, Sheryl and Paula-Ann Cech (1996). "Subject Pool Choice and Treatment Effects in Economic Laboratory Research." *Research in Experimental Economics*, 6, 239–292.

Benabou, Roland and Jean Tirole (2002). "Self-Confidence and Personal Motivation." *Quarterly Journal of Economics*, 117, 871–915.

Berg, Joyce, John W. Dickhaut, and Kevin A. McCabe (1995). "Trust, Reciprocity, and Social History." *Games and Economic Behavior,* 10, 122–142.

Bewley, Truman F. (1995). "A Depressed Labor Market as Explained by Participants." *American Economic Review*, *Papers and Proceedings,* 85(2), 250–254.

Bewley, Truman F. (1999). *Why Wages Don't Fall During a Recession*. Harvard University Press.

Bohnet, Iris, Bruno S. Frey, and Steffen Huck (2001). "More Order with Less Law: On Contract Enforcement, Trust, and Crowding." *American Political Science Review,* 95(1), 131–144.

Bolton, Gary E. and Axel Ockenfels (2000). "ERC—A Theory of Equity, Reciprocity, and Competition." *American Economic Review,* 90(1), 166–193.

Camerer, Colin (2003). *Behavioral Game Theory*. Princeton University Press.

Camerer, Colin, Teck Ho, and Kuan Chong (2003). "Models of Thinking, Learning, and Teaching in Games." *American Economic Review*, *Papers and Proceedings*, 93(2), 192–195.

Camerer, C., Teck-Hua Ho, and Juin-Kuan Chong (2004). "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics*, forthcoming.

Cameron, Judy and W. David Pierce, (1994). "Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis." *Review of Educational Research,* 64, 363–423.

Chiappori, Pierre-Andre and Bernard Salanie (2003). "Testing Contract Theory: A Survey of Some Recent Work." In: *Advances in Economic Theory, Eighth World Congress of the Econometric Society* by Mathias Dewatripont, Lars Peter Hansen, Stephen J. Turnovski. Cambridge University Press.

Cooper, David J., John H. Kagel, W. Lo, and Liang Gu Qing (1999). "Gaming Against Managers in Incentive Systems: Experiments with Chinese Students and Chinese Managers." *American Economic Review*, 89, 781–804.

Deci, Edward L. (1971). "The Effects of Externally Mediated Rewards on Intrinsic Motivation." *Journal of Personality and Social Psychology,* 18, 105–115.

Deci, Edward L., R. M. Koestner, and R. Ryan (1999). "A Meta-Analytic Review of Experiments Examining the Effect of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin*, 125, 627–668.

Dufwenberg, Martin and Georg Kirchsteiger (2004). "A Theory of Sequential Reciprocity." *Games and Economic Behavior*, 47, 268–298.

Falk, Armin and Urs Fischbacher (1999). "A Theory of Reciprocity." Working Paper No. 6, Institute for Empirical Research in Economics, University of Zürich.

Fehr, Ernst and Simon Gächter (2002). "Do Incentive Contracts Undermine Voluntary Cooperation?" Working Paper No. 34, Institute for Empirical Research in Economics, University of Zürich.

Fehr, Ernst and Bettina Rockenbach (2003). "Detrimental Effects of Sanctions on Human Altruism." *Nature,* 422, 137–140.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl (1993). "Does Fairness Prevent Market Clearing? An Experimental Investigation." *Quarterly Journal of Economics,* 58, 437–460.

Fehr, Ernst and Klaus M. Schmidt (1999). "A Theory of Fairness, Competition and Cooperation." *Quarterly Journal of Economics,* 114, 817–868.

Fehr, Ernst and Klaus M. Schmidt (2003). "Theories of Fairness and Reciprocity—Evidence and Economic Applications." In *Advances in Economics and Econometrics*, edited by M. Dewatripont, L. Hansen, and S. J. Turnovsky. Cambridge University Press.

Frey, Bruno S. and Felix Oberholzer-Gee (1997). "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding Out." *American Economic Review,* 87, 746–755.

Gneezy, Uri and Aldo Rustichini (2000a). "A Fine is a Price." *Journal of Legal Studies,* 29, 1–17.

Gneezy, Uri and Aldo Rustichini (2000b). "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics,* 115(2), 791–810.

Haigh, Michael S. and John A. List (2004). "Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis." *Journal of Finance,* forthcoming.

Hannan, R. Lynn, John H. Kagel, and Donald V. Moser (2002). "Partial Gift Exchange in an Experimental Labor Market: Impact of Subject Population Differences, Productivity Differences, and Effort Requests on Behavior." *Journal of Labor Economics,* 20(4), 923–951.

Heath, Chip (1999). "On the Social Psychology of Agency Relationships: Lay Theories of Motivation Overemphasize Extrinsic Incentives." *Organizational Behavior and Human Decision Processes,* 78, 25–62.

Kreps, David M. (1997). "Intrinsic Motivation and Extrinsic Incentives." *American Economic Review,* 87, 359–364.

Macaulay, Stewart (1963). "Non-Contractual Relations in Business: A Preliminary Study." *American Sociological Review,* 28, 55–70.

Prendergast, Canice (1999). "The Provision of Incentives in Firms." *Journal of Economic Literature,* 37, 7–63.

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

Sitkin, Sim B. and Nancy L. Roth (1993). "Explaining the Limited Effectiveness of Legalistic Remedies for Trust/Distrust." *Organization Science,* 4, 367–394.

Sobel, Joel (2002). "Social Preferences and Reciprocity," Working paper, Department of Economics, University of California, San Diego.