WHEN PUNISHMENT FAILS:
RESEARCH ON SANCTIONS, INTENTIONS AND NON-COOPERATION

BY

DANIEL HOUSER ERTE XIAO KEVIN MCCABE VERNON SMITH

INTERDISCIPLINARY CENTER FOR ECONOMIC SCIENCE
GEORGE MASON UNIVERSITY

February 17, 2005

Comments Welcome
dhouser@gmu.edu exiao@gmu.edu

Abstract: People can become less cooperative when threatened with sanctions, and researchers have pointed to both "intentions" and incentives as sources of this effect. This paper reports data from a novel experimental design aimed at determining the relative importance of intentions and incentives in producing non-cooperative behavior in a personal exchange environment. Subjects play one-shot investment games in pairs. Investors send an amount to trustees and request a return on this investment and, in some treatments, are given the option to threaten sanctions to enforce this return request. The decisions of trustees who face credible threats intentionally imposed (or not) by their investors are compared to the decisions of trustees who face threats randomly imposed (or not) by nature. When not threatened, trustees typically decide to return a positive amount that is less than the investor requested. When threatened with sanctions this decision becomes least common. In particular, under severe sanction threats most trustees return the desired amount, while under weak threats the most common decision is to return nothing. Critically, these results do not depend on whether the trustee is threatened intentionally by their investor or randomly by nature: trustees who are threatened with weak sanctions are significantly more likely to provide a zero return to their investors, even when they know that their investors had no role in imposing the threat. Our findings lend support to the view that credible threats of sanctions generate a "cognitive shift" that crowds-out norm-based motivations and increases the likelihood of income-maximizing behavior.

Acknowledgements: We thank the International Foundation for Research in Experimental Economics and the National Science Foundation (SES-0129744 and SES-0339181) for funding that supported this research. We are grateful to Ernst Fehr and Bettina Rockenbach for sending us data and instructions related to their 2003 paper. We received useful comments on earlier drafts of this paper from James Andreoni, Colin Camerer, Tyler Cowen, John Dickhaut, David Dickinson, Martin Dufwenberg, Read Montague, Francesco Parisi, Aldo Rustichini and Bart Wilson, as well as participants at the 2004 SEA meetings, and seminars at Max Plank (Bonn), Vasser College, University of Mannheim, NIH/NIDA, the Interdisciplinary Center for Economic Science and George Mason University's graduate student workshops. We thank ICES graduate students, particularly Bridget Butkevich, for assistance with experiments.

I. Introduction

Sanctions are often written into incomplete contracts in an effort to encourage cooperation. Especially when information is asymmetric, the threat of sanctions might discourage self-interested actors from cheating or making other opportunistic decisions. In many cases, acts that constitute "cooperation" are explicitly described. For example, legislation might require firms to reduce effluents to a certain level, and stipulate that not doing so would result in a known monetary forfeiture. The threat might be successful in that firms might cooperate and reduce effluents as directed. Alternatively, the threat might fail in any number of ways. For example, firms might reduce effluents but still produce more than directed. Or firms might make no reduction efforts at all, or even produce more than they had previously. This paper uses a novel experiment to examine the conditions under which credible threats of sanctions are likely to promote cooperation and, when not, how they are likely to fail.

A large literature in both psychology and economics makes clear that "intentions" and "incentives" are two key factors in determining sanctions' behavioral effects (see, e.g., Yamagishi, 1986, 1988; Ostrom, et al., 1992; Fehr and Falk, 2002; Fehr and Rockenbach, 2003; Fehr and List, 2004; Dickinson, 2001; Andreoni, et al, 2003; Bewley, 1999; Gneezy and Rustichini, 2000a). In general, intentions and incentives are distinguished by the fact that intentions involve personalized rules by men, while incentives involve impersonal rules by law. Here, "incentives" will refer narrowly to the pecuniary tradeoffs a sanctioning mechanism creates. "Intention" effects will be those that stem from a trustee's belief regarding an investor's motivation for threatening (or not threatening) a sanction. For example, a trustee might become angry at an investor who makes a threat if he interprets it as a signal of mistrust, and consequently act in a way that does not maximize his or her earnings.

Although it is widely agreed that both intentions and incentives can matter (e.g., Camerer, 2003, highlights the potential importance of intentions in a wide variety of game-theoretic contexts), there has not previously appeared any systematic evidence on their relative importance in enforcing cooperation through punishment in trust environments. Given the ubiquitous use of sanction threats in such environments (e.g., employer and employee, or principal and agent), a deeper understanding of the sources of

their behavioral consequences is evidently important. We present here a novel experimental design to investigate the relative behavioral effects of punishment intentions and punishment mechanism incentives in a gift exchange context. Moreover, our design allows us to distinguish "types" of non-cooperative behavior, and we draw inferences with respect to the way non-cooperative decision making differs among various incentive and intention conditions.

We extend an experimental design used in a study of sanctions by Fehr and Rockenbach (2003); (see also Fehr and List, 2004). Broadly, this design is an "investment" game (Berg, Dickhaut and McCabe, 1995) where an investor sends an amount to a trustee, that amount is tripled by the experimenter, and then the trustee sends some fraction of the tripled amount back to the investor. Like Fehr and Rockenbach (2003), we augment this environment by allowing investors to threaten trustees credibly with sanctions if an insufficient amount is back-transferred to the investor. Our design is novel in that we compare outcomes in that "intentions" treatment to a new treatment where the sanction is randomly assigned to trustees by nature. In this way are able to separate cleanly intention effects from incentive effects of sanctions.

Our design also distinguishes between "weak" and "severe" punishment incentives. In particular, the ratio of the sanction to the back-transfer request is a natural measure of the sanction's severity. When the threatened sanction is "large" relative to the back-transfer request we say that the sanction is severe, and otherwise we say it is weak. We obtain observations on both cases by holding exogenously fixed the sanction's amount, while the back-transfer request varies according to first-movers decisions. In this way we obtain substantial variation in the severity of the threatened sanction, and this variation identifies the effect of the incentive's severity on trustees' decisions.

We focus almost entirely on cases where the investor requests a "fair" backtransfer, where by fair we mean that investor requested at most 2/3 of the tripled transfer amount. (If 2/3 of the tripled amount is returned, then both the investor and trustee earn the same amount in the experiment.) One reason is that many policy relevant sanctioning schemes are at least perceived as enforcing fair outcomes, so knowing more about this case is important. Moreover, as a practical matter, most of the investors in our design requested fair back-transfers. Consequently, collecting enough data to appropriately control for the

"unfair" intention effect, and separate it from the punishment intention effect, would be highly burdensome. At the same time, the unfair effect is clearly important: we find that trustees behave statistically significantly differently when faced with an unfair backtransfer requests than they do otherwise.

Our design turns out to have surprising power: we both clarify and significantly extend Fehr and Rockenbach (2003) and others' related findings. When not threatened, trustees typically decide to return a positive amount that is less than the investor requested. When threatened with sanctions this decision becomes least common. In particular, under severe sanction threats most trustees return the desired amount, while under weak threats the most common decision is to return nothing.[1] Moreover, we were surprised to find that these results hold both when the sanction is threatened intentionally by the investor as well as when it is threatened randomly by nature: a trustee is more likely to return nothing to his or her investor after having been threatened with a weak sanction, regardless of whether the investor played any role in determining that sanction.

This paper is in five sections. The next gives additional background on research in psychology and economics on sanctions. The third and fourth sections detail our design and results, respectively. The fifth section is a concluding discussion.

II. Background

In this section we describe some of what is known about the way incentives and intentions affect the efficacy of sanctions. The literature in this area is vast: see Fehr and Falk (2002) for a comprehensive survey of this topic. We also briefly discuss previous research on non-cooperative decision making. We point out that, to the best of our knowledge, ours is the first attempt to compare and contrast the distributions of "types" of non-cooperative decisions under various incentive and intention structures.

[1] Other studies have found similar all-or-nothing behavior (see, e.g., Dickinson, 2001; Fehr and Gächter, 2002; Tyran and Feld, 2004). However, we are not aware of any study that distinguishes the relative importance of intention from incentive effects in these all-or-nothing decisions, nor have these studies recognized the significance of weak and severe punishment in affecting the distribution of returns.

II.1 Incentive Effects

It is useful to think of incentive effects as stemming from a sanction threat imposed by nature, rather than by another person. Standard economic arguments show that incentives can be effective in enforcing cooperation. In particular, threats of sanctions can make a non-cooperative action's expected net benefit negative, so that income-maximizing agents will not take the action. Many have also pointed out that incentives can reduce cooperation (see, e.g. Kreps, 1997, Frey and Oberholzer-Gee, 1997, Benabou and Tirole, 2003, Fehr and Falk, 2002). An important example is Gneezy and Rustichini (2000b), who report that when subjects were offered small monetary incentives to perform a task, they performed more poorly than those who were offered no compensation (but not as well as those who were offered more substantial compensation). These results, they argue, suggest that monetary rewards can interfere with an intrinsic desire to perform a task well.

This line of reasoning is familiar to psychologists, who have long argued that pecuniary incentives can make people less interested in desirable conduct for its own sake (see. e.g. Deci, et al, 1999; Lepper and Greene, 1978). The underlying idea is that individuals are likely to attribute their actions to intrinsic motivation when there is no external incentive present, but to discount this motive when a pecuniary reward is offered for actions. Tenbrunsel and Messick (1999) proceed in a similar spirit, and argue that reward and sanctioning systems can increase the likelihood that a cognitive "business" frame, rather than an "ethical" frame, will be used to make decisions.

Overall then, previous research on incentives reveals that sanctioning mechanisms can increase cooperation by changing the payoff structure, but also include the downside risk that norm-based cooperation can be decreased due to a shift in the cognitive decision frame. This has several relevant implications for this paper. First, when a threatened sanction is severe (in the sense that the cost of the sanction greatly exceeds the benefit of non-cooperation), the incentive effect implies that people will avoid punishment and cooperate perfectly. Second, when the threatened sanction is weak, the incentive-driven cognitive-shift will lead people to take the action that maximizes their own earnings, even if it is highly non-cooperative. Finally, in the absence of threat-based incentives, subjects will make their decision within more of an ethical context and complete defection will be less frequent than in the weak sanction case.

II. 2. Intention Effects

When a credible threat is endogenous, in the sense that one person chooses to threaten another, both the incentives of the mechanism and also the intentions that underlie the threat can affect behavior. Indeed, humans seem strongly disposed to infer intentionality when understanding others' actions (see e.g., Gibbs, 1999). For example, imposing sanctions can be seen as a signal of distrust (see, e.g. Fehr and Falk, 2002), or might create a hostile atmosphere (Bewley, 1999), and consequently reduce cooperation. Similarly, intentional acts of helping are more likely to be reciprocated than unintentional ones, and intentional acts of aggression are responded to more often than unintentional acts (see, e.g., Greenberg and Frisch, 1972; Gordon and Bowlby, 1989; Blount, 1995; Offerman, 2002).

Many economic models now include intention effects. Rabin (1993) is an early approach to incorporating the perceived kindness of another into one's own preference structure. Another nice example is Dufwenberg and Kirchsteiger (1998), who develop a theory of reciprocity for extensive form games. A substantial amount of experimental research also suggests that intentions can play an important role in shaping decisions (see e.g., McCabe, Rigdon and Smith, 2003; Fehr, Gächter, 2000; Nelson, 2002; and Charness, 2002).[2]

Previous theoretical and experimental research suggests that trustees' decisions in our experiment could be affected by their perception of investors' intentions. For example, if investors who threaten sanctions are perceived as being unkind or distrusting[3], then the intention effect will lead to less trustee cooperation in this case than when otherwise identical sanctions are threatened by nature. Moreover, intentions might be particularly salient, and lead to relatively less cooperation than when punishment is not intentionally imposed, in cases where an investor threatens severe sanctions. Note this stands in sharp contrast to the incentive effect of severe sanctions. Intentions, of course, can also have a positive effect on cooperation. In particular, investors who choose not to

[2] Bolton, et al. (1998) report data from a design where intentions do not seem to affect behavior.
[3] Charness and Dufwenberg (2004) suggest an alternative source of intention effects. Loosely speaking, their argument is that investors who threaten sanctions might be perceived by trustees as expecting a low return, and that trustees would therefore feel less guilty about providing a low return.

sanction might be perceived as "nice" and be relatively highly rewarded for this. Our experimental design provides transparent inference with respect to the way intentions influence the reaction to threats, or non-threats, of both severe and weak sanctions.

II.3. Non-cooperative Decision Making

Credible threats of sanctions, both in the laboratory and naturally occurring world, typically stipulate that a particular action is cooperative, while any other action is non-cooperative and subject to sanctions. In the event that a person chooses not to cooperate under such a threat, what sort of non-cooperative decision is likely to be made? How do non-cooperative decisions under threats compare to non-cooperative decisions when no threats are present? Do the answers to these questions depend on whether the threat is intentional? Our design sheds light on these questions.

Much experimental research on cooperation uses extensive form games that restrict subjects' action spaces to two alternatives: one cooperative and one not (see, McCabe, Rigdon and Smith, 2003). Clearly, while such designs have the twin advantages of being clean and providing relatively easily interpretable results, they also cannot inform the above questions. Moreover, even when the design can potentially inform the nature of non-cooperation, studies have typically not included the randomization treatments necessary to inform intention effects (see, e.g., Fehr et. al. 1997).

The important paper by Fehr and Rockenbach (2003) is an example of this point. Although their design is in principle capable of distinguishing types of non-cooperative decisions under different incentive structures, it is not their paper's point to do so. In fact, there are at least four natural "types" of decisions that are available to trustees in their design. They can (i) send nothing back (investors lose all the endowment); (ii) send back a positive amount less than the transfer amount (investors lose some of the endowment); (iii) send back at least the transfer amount but less than the requested amount (investors earn at least the original endowment); (iv) send back at least the requested amount (investors get at least what they want). Note that the first three are non-cooperative in different ways, while the last is cooperative. In this paper we provide new evidence on

the way the distributions of types of cooperative and non-cooperative decisions are influenced by sanctions with various incentive and intention structures.[4]

III. Experiment Design

Because our experiment is closely connected to Fehr and Rockenbach (2003) (henceforth, FR), we begin this section with a discussion of their procedures. Doing this allows us to distinguish our work from theirs, and to indicate how our methods substantially sharpen our understanding of the behaviors observed in both their environment and our own.

III.1. Fehr and Rockenbach's Study

FR study sanction effects in a modified investment game (Berg, Dickhaut and McCabe, 1995). In their "trust" treatment, both the investor and the trustee receive an endowment of ten money units (MUs). The investor sends some, all or none of his endowment to the trustee, and the experimenter triples any amount sent. In addition, the investor specifies the amount, between zero and the entire tripled amount, that she would like the trustee to return. After seeing the tripled amount sent and the desired back-transfer, the trustee sends some, all or none of the tripled amount back to the investor. The investor earns his endowment of 10 MUs, minus anything transferred to the trustee, plus any back-transfer amount. The trustee earns the endowment of 10 MUs, plus the tripled transfer amount, minus any amount back-transferred.

FR study behavior in a second treatment, the "incentive" condition, which is identical to the trust baseline except that the investor can now choose whether to commit to imposing a fine of a fixed four MUs on the trustee if less than the desired amount is returned. Both subjects are aware that this fine, if due, does not accrue to the investor but rather to the experimenter's research budget. When the trustee makes his decision in the incentive condition he knows which sanction option his investor has chosen.

FR's major finding is that, on average, trustees' back-transfers were highest when the investor voluntarily refrained from the fine in the incentive condition, and lowest

[4] The fact that we study a one-shot environment of course limits the amount of type classification that one can do. Cooperative types have been studied in dynamic environments by Houser and Kurzban (2002) and Kurzban and Houser (2004), among others. A robust statistical procedure for behavioral type classification has been provided by Houser et. al. (2004).

when the investor imposed the fine. The mean back-transfer in the trust condition fell between the two means in the incentive condition. Hence, FR provided evidence that the use of sanctions can reduce cooperative behavior.[5]

FR's experiment incorporates both the intention to punish and a punishment incentive mechanism, but their design does not distinguish these two effects. Consequently, one cannot know how much of the reduction in cooperation is due to the punishment incentive itself, and what amount is due to the intention to use this incentive.[6]

III.2 Design

Our design enables us to examine the effect of sanctions when they are assigned randomly to trustees, and in a way that eliminates investor intention effects. Intuitively, comparing the condition without punishment intentions to the condition with punishment intentions enables inferences regarding punishment intention effects.

Treatments

Figure A describes our "intentions" treatment. Note that it corresponds exactly to FR's "incentive" condition.

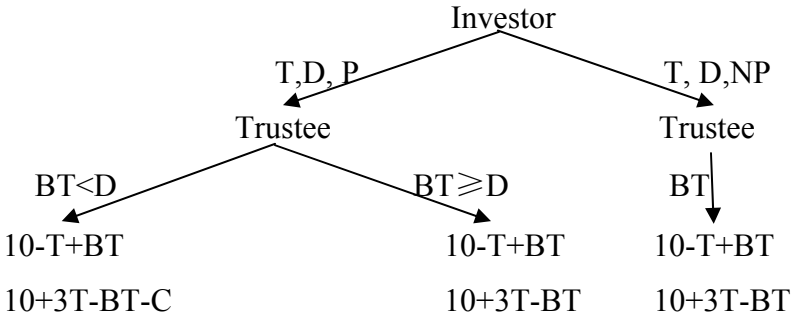


Figure A: Intention treatment

T—investor's transfer; D—investor's desired back transfer; P—threaten punishment; NP—not threaten punishment; BT—Trustee's backtransfer; C—amount of fine (payoff cut) if BT is less than D.

[5] Fehr and List (2004) find similar results using the same design with a novel subject pool.

[6] Note also that in FR's data it turns out that investors who choose to fine also send less, on average, than those who choose not to fine. Because all investors start with the same endowment, this means that investors who fine send a smaller percentage of their endowment than those who don't. FR's main result is that if the fine was imposed the trustees paid back a smaller percentage of the tripled investment than if the fine was not imposed (30.3% as compared to 47.6%.) From FR's design, we are unable to know how much of this difference might be unrelated to incentive or intention effects, and rather simply due to differences in the percentage of the investors' endowments' received by trustees in the two cases. We control for this.

In this treatment subjects are paired with each other anonymously. Both investors and trustees are endowed with $10 at the beginning of the experiment. The investor first decides how much to transfer to her trustee (T), how much to request as a backtransfer (D), and whether to threaten punishment (P or NP). With this information in hand, the trustee then decides how much to transfer back to the investor. If the trustee returns less than the investor requested, and if the investor chose to impose the conditional payoff cut, then the final earnings of the trustee are reduced by $4, and this amount is known by both the investor and the trustee. Both the trustee and investor know that, if it is due, this $4 sanction does not go to the investor, but instead remains in the experimenter's research budget.

Our second treatment, the "random" treatment, is described by Figure B.

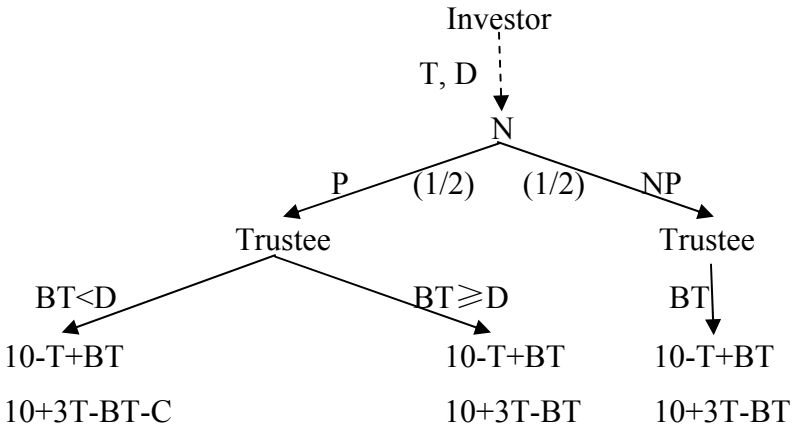


Figure B. random treatment

In the "random" treatment, whether a given trustee is subject to a sanction threat is determined through a transparent randomization procedure. We ran each of these sessions with multiples of four subjects, which allowed us to ensure that exactly half of the trustees could be randomly assigned to the punishment condition, and the other half to the non-punishment condition. The assignment of subjects to conditions was performed in front of the subjects, and involved simply drawing envelopes from a box.

The goal of our random treatment was to eliminate any possibility that trustees could believe investors had punishment intentions. Our approach to this was to give investors in the random treatment absolutely no information about sanctions that the

trustees might face, and to tell trustees that this is the case (this is indicated by the dotted line in Fig. B). We believe that blinding investors to the possibility that the trustee will be sanctioned is necessary for clean and compelling separation of intention and incentive effects. In particular, if investors knew there was a possibility that trustees could be sanctioned, then trustees who are randomly assigned to the sanction condition might believe (perhaps appropriately) that their investors "expected" or "hoped" that would happen. Taking this view might be particularly likely if, as sometimes happens, the investor sent one dollar and asked for all three dollars back. It seems likely that trustees' decisions could be affected by such an intention attribution, and this would act to confound (to an unknown extent) our ability to separate intention and incentive effects.

An obvious consequence of blinding investors to the sanctioning procedures is that, while they are of course not deceived, they also do not have full information about the way trustees' payoffs are determined. This has the usual consequence that we lose control over what investors believe regarding trustees' earnings and the way that might affect their own earnings. Accordingly, interpreting investor behavior becomes difficult. *Consequently*, *we draw no inferences regarding motivations for investor behavior. Our conclusions are based entirely on the behaviors of trustees.*

Although investors in the random treatment are not aware of the sanctions at the time they make their decisions, the sanctioning procedure is explained to them at the end of each session, and at that time they are also informed of the condition to which their matched trustee was randomly assigned. This is done to ensure maximal symmetry with the intention treatment. In particular, trustees in the intention treatment know that their investors know whether they have been assigned to the sanction condition. So, in the random treatment trustees are informed that the investors will be told about the sanctioning procedure and their matched trustee's assignment at the session's end.

An attempt at symmetry is also the reason that we did not run "sanction everybody" or "sanction nobody" treatments, which is a common suggestion offered to us when presenting this paper. Note that in the intention treatment there are three pieces of information simultaneously revealed to trustees: the transfer amount, the desired back-transfer and the punishment condition. There is uncertainty about all three pieces of information up to the point that the investors' decisions are revealed to them. In contrast,

in a "sanction everybody" treatment, for example, all trustees would know from the beginning that they will face the sanction incentive. They will learn only two new pieces of information directly prior to making their decision. The extent to which this might matter is an empirical question; to us the important point is simply that it might. Although we conduct the randomization in a transparent way in front of the subjects, the assignments generated by the randomization are revealed to them at the same time, and in the same way, as occurs in the incentive treatment.

To summarize, the salient difference between our random and intention treatments is that investors cannot have any punishment intentions in the former, and it is not reasonable to expect that trustees might assign punishment intentions to them. Consequently, any differences in the behaviors of trustees between these two treatments can be cleanly attributed to differences in investors' sanction intentions. Moreover, as there are no punishment intentions in the random treatment, we can study the punishment incentive effect *within* this treatment by examining how the trustees in the punishment group behave in relation to those trustees assigned to the non-punishment group.

III.3 Procedures

A total of 532 subjects participated in this experiment. 149 pairs played in intention treatment and 117 pairs played in random treatment[7]. All subjects were recruited from George Mason University's general student population, using standard recruiting procedures in place at the Interdisciplinary Center for Economic Science. Subjects earned a $5 show up bonus for arriving to the lab on time. Subjects earned E$ during the experiment, and at the end of the experiment the E$ were exchanged for dollars at the rate of 1 to 1.

Our specific procedures are detailed in the instructions found in Appendix B. Key aspects of our procedures include the following. Each treatment included just one game (one decision by the investor and one decision by the trustee), and subjects knew that this would be the case. All treatments were run by hand, were single blind and began with each subject being randomly assigned to one of two roles, "investor" or "trustee."

[7] In the random treatment, there were three investors who transferred zero and consequently, asked for zero back. Because the trustee has no decision to make in these cases, we drop these data from our analysis, leaving 114 observations in the random treatment.

Investors (called Actor 1 in the instructions to avoid context effects) were led to one room and trustees (Actor 2 in the instructions) another. After they were separated each trustee was randomly and anonymously matched with one investor. The investors wrote their decisions on cards that were then delivered by the experimenter to the trustees. The trustees wrote down their decisions on the same cards, and those decisions were delivered back to the investor by the experimenter. Earnings were calculated, subjects were paid and the experiment concluded. On average, subjects were in the lab for about 90 minutes and earned about $15 in addition to the show-up bonus.

It is worthwhile to reiterate the procedures we used to eliminate investors' punishment intentions. In the random treatment, investors were not informed that some trustees would be threatened with sanctions until after the trustees' decisions were complete, and trustees were made aware that this was the case. Also, the randomization procedure used to assign trustees to the sanction condition was highly transparent. It involved blindly choosing envelopes out of a box, in front of the trustees, and assigning the sanction condition based on that draw. Trustees were attentive to the randomization process. They clearly understood that the sanction condition to which they were assigned was determined using a fair randomization procedure controlled by the experimenter, and that investors were not at all connected to that procedure.

IV. Results

IV.1 The Data

Table 1 provides aggregate information on the decisions of investors and trustees. Panel A describes the results for the cases with "fair" backtransfer requests, where we define fair to mean any request that is less than or equal to 2/3 of the tripled amount.[8] Panel B provides information on the relatively small number of cases where the request was

[8] The distribution of investor decisions should be interpreted with caution. Following a treatment reported by Fehr and Rockenbach (2003), we provided investors (not trustees) in all of our treatments with a graph describing the FR (2003) results (of course excluding results related to punishment in the "random" treatments.) Our goal was to study behavior under "fair" back-transfer requests, and we conjectured that providing (only) investors with the FR (2003) results could act to increase the likelihood of fair requests. While FR (2003) found no statistically significant effect of their information sheet, we do not have the contrast available to assess its impact.

unfair. The fair and unfair data sets are statistically significantly different, especially when punishment is imposed, and cannot be pooled. This is easily seen by comparing, for example, the percent of the tripled transfer amount that trustees return. When the request is fair and a sanction is randomly imposed about 32% is returned, while in the unfair case about 12% is returned, and the difference is statistically significant ($p<0.05$). Our main interest in this paper is in fair backtransfer requests and we focus the first three parts of this section, and most of our conclusions, on those data. We briefly discuss the unfair backtransfer requests at the end of this section.

Table 1(A) shows there are a total of 96 trustees in the incentive treatment who receive a fair backtransfer request, and 44 of these trustees are randomly assigned to the sanction condition. In the intention treatment, 122 trustees were asked to return less than or equal to the equal-split amount, of whom 62 face the conditional payoff cut. The aggregate data does not suggest much difference between the intention and incentive treatments. In both cases, under threats of punishment trustees return about 32% of the tripled investment amount.

To investigate whether there are effects stemming from a sanction's severity we say a sanction threat is "severe" when an investors' request is less than 8E$, and we say that a threatened sanction is "weak" otherwise. In the random treatment, about 42% of trustees who do not face threats of sanctions are asked to return less than 8E$, while that number is 45% for trustees assigned to the conditional payoff cut. In the intention treatment those same numbers are 42% and 47% for the no-threat and threat cases, respectively.

IV.2. Intentions, Incentives and Return Amounts

To learn about sanction intention effects, we must control for other potential effects on trustees' decisions. The trustee receives three messages: a transfer amount, a desired back-transfer and whether a sanction has been threatened (either by the investor or nature). Our strategy is to model trustees' decisions as a function of those three messages and an error component. In all cases, we assume the error component is independent across subjects in different experimental sessions, but potentially correlated among subjects within the same experimental session.

We begin by investigating the way in which these signals affect the percentage of the tripled amount (sent by the investor) that is sent back by the trustee. Using percentages provides a first control for the effect of the transferred amount. The percentage amount returned can itself, however, be affected by the transfer amount. As we argued in section III.1 above, trustees might "percentage match", in which case the percentage of the tripled amount that they return would increase in the percentage of the investor's endowment that was sent. Our analysis controls for this effect by including a dummy for whether the investor's transfer was high (four or more). We chose four because it is equal to the amount of the sanction. Our results are robust to changes in that number, or to replacing the dummy with the actual transfer amount.

It is easy to see that the effect of the sanction might depend on the amount of the desired backtransfer. Consider, for example, the case where an investor sends two, desires a backtransfer of three, and chooses to punish (i.e. punishment is severe). A trustee who returns less than three is sanctioned by four, and consequently has a very strong incentive to cooperate. On the other hand, an investor who sends ten, desires a backtransfer of 20 and threatens a (weak) sanction places the trustee in a very different situation. In this case a completely noncooperative trustee, i.e. return nothing, will earn 16 more than a trustee who returns the investor's desired amount. We control for this effect by interacting dummies that specify whether punishment was chosen intentionally or randomly, and whether the desired backtransfer amount was eight or more (high request), or less than eight (low request). The use of eight as the cutoff point is obviously arbitrary, but was guided by the fact that eight is double the sanction amount. Intuitively, the cost of cooperation might begin to seem large when the amount that must be returned is more than twice the sanction amount. Our results are robust to smaller or larger cutoff values, and can be obtained from the authors on request.

Table 2 provides the results of an OLS regression (allowing for correlated error-components within sessions) of the percentage of tripled amount returned on the regressors described above. In particular, the nine regressors include a dummy for whether the investment amount was high (greater than four) and eight terms determined by interacting decisions to punish and not punish, randomly or intentionally, with the high and low request conditions. We report OLS results because they are easy to interpret.

The results are not substantively changed by running a Tobit that accounts for the fact that returned amounts are censored at 0% and 100% (in fact there are no cases where the entire tripled amount is returned).

From Table 2 one sees that the R-squared for our simple specification is 0.59, suggesting that our model provides a reasonable fit to the data. The coefficient of the transfer amount dummy is unexpectedly negative, but rather small and statistically insignificant. This suggests that transfer effects are adequately controlled by using percentage of tripled amount returned as the dependent variable. Note that all of the coefficients involving punishment in Table 2 are positive.

Casual observation of the coefficient estimates in Table 2 reveals that coefficients on four analogous random and intention conditions tend to be quite similar. A straightforward way to provide evidence on punishment intention effects is to test jointly the four restrictions that each intention coefficient is equal to its corresponding random coefficient. That is, the null hypothesis is that the coefficient on intention punish is equal to the coefficient on random punish, and the coefficient on intention no punish is equal to that on random no punish, and so on. The alternative hypothesis is that the coefficients in at least one pair take different values. This test is unable to reject the null hypothesis ($F(4,35)=0.22$, p-value=0.93). This provides strong evidence that punishment threatened by investors does not affect trustees' returns differently than threats imposed by nature.

On the other hand, an analogous test for the effect of punishment (that is, random punishment equals random no punishment, and so on for each of the three remaining pairs) provides some evidence against the null hypothesis that mean returns by trustees are not affected by the presence of punishment ($F(4,35)=2.23$, p-value=0.09). Moreover, the results of the test are stronger if we restrict attention to only the two pairs of low request coefficients ($F(2,35)=4.36$, p-value=0.02). This provides evidence that threats of sanctions interact with the request amount to influence trustees' decisions.

Figure 1 summarizes these findings by plotting the percentage of tripled amount returned against the amount sent for each of the eight conditions of interest (e.g., random threatened high request, intention threatened high request, and so on.) The legend to Figure 1 also details the number of trustees observed in each of these eight cases. It is clear from casual observation that neither the transfer amount nor intentions have much

of an affect on mean returns. On the other hand, there is evidence of sanction effects, and evidence supporting an interaction effect between the sanction and the request.

Because we have found no evidence of intention effects, we proceed with our analysis of return decisions by pooling the random and intention data, so that the eight cells analyzed above collapse to four. The resulting data are described in Figures 2 and 3, which provide greater detail on the way threats influence trustees' returns. Consider first Figure 2. This figure provides a histogram of the percentage of tripled amount returned, in the threat and no threat conditions, when the desired backtransfer request is low (less than eight). The average percent of tripled amount returned is higher when the trustee is threatened with punishment (43.4 vs. 26.0, p=0.002). More striking, however, is that the distribution under no threat as rather flat, while the distribution under punishment threats has a "U" shape. A Kolmogorov-Smirnov test finds no support for the null hypothesis that the distributions are identical (p=0.0). A reason is that the two distributions have statistically significantly different standard deviations: 30.4 under punishment threats as compared to 22.7 when punishment is not threatened ($F(49,47)=1.79$, p=0.02).

Figure 3 provides the same type of information for the two high request cells. In this case the mean of the punishment threats distribution (22.0) lies below the mean of the no punish threats distribution (29.5), although the difference is not statistically significant. Again, however, a Kolmogorov-Smirnov test finds evidence that the distributions are not the same (p=0.01) and again this is reflected in differences between the two distributions' standard deviations. Under threats of punishment, the standard deviation of 27.8, while it is 23.6 when punishment is not threatened ($F(57,65)=1.38$, p=0.10).

In the next section we examine the source of the spread in returns that occurs when punishment is threatened. We specify and estimate a multinomial choice model in order to provide evidence that, when threatened with punishment, trustees tend either to return the entire amount that the investor requested or to return nothing. Said another way, trustees are not likely to return only some of the desired amount when they have been threatened with sanctions, and this is true regardless of whether the threat is intentional.

IV.3. Sanctions and Non-Cooperation

Investors threaten trustees with sanctions in an effort to encourage them to return the desired back transfer. We call trustees "cooperative" if they return at least the requested amount, and non-cooperative otherwise. We use this definition because we analyze only those cases where the investor requested at most half of the pair's aggregate earnings. To provide less than an equal split, when an equal split was explicitly requested, is reasonably viewed as (at least) weakly non-cooperative. In this section we present the results from two multinomial choice models of trustee decision making. The first model, which includes a relatively course action space but a rich set of explanatory variables, allows inference with respect to how a trustee's decision to send some, all or none of the desired back transfer is affected by intentionally or randomly imposed sanctions. As above, we find that trustees' decisions are statistically identical in both the intention and random conditions. Consequently, we estimate a second model that eliminates the intention variables but refines the action space.

We begin by defining a trustee's choice set so that it includes three mutually exclusive and exhaustive alternatives: cooperate completely (send at least as much as the investor requested) and two non-cooperative options: defect completely (send nothing), or weakly non-cooperate by sending a positive amount that is less than the requested amount. As described formally in Appendix A, we adopt a random utility specification to analyze these data.[9] This specification implies that the utility associated with each alternative depends on the nine explanatory variables discussed above and an alternative and subject specific error component which might be correlated among individuals within the same experimental session. We assume that each subject chooses the alternative associated with his or her highest subjective utility value. We assume a structure for the error components that implies choices can be characterized by a multinomial logit.

IV.3.a. Three-alternative model

Table 3 presents the multinomial logit estimates, where weak non-cooperation is taken as the baseline. Multinomial logit coefficients are difficult to interpret, but casual inspection of their values quickly reveals that the transfer amount is not statistically significant in

[9] We point out in Appendix A that it is inappropriate to use an ordered specification to analyze this data (e.g., ordered logit or ordered probit).

either equation, and analogous random and intention coefficient estimates are similar relative to the standard errors of the estimates. In fact, a joint test of the null hypothesis that, across the eight pairs of analogous terms, the random and intention coefficients are identical cannot be rejected at standard significance levels (chisq(8)=6.1, p=0.63). This is convergent evidence that intentional threats by investors do not change trustees' decisions in comparison to cases where a sanction is threatened randomly by nature.

On the other hand, a joint test of sanction effects (e.g., random punish = random no punish, and so on) clearly indicates that trustees' decisions to cooperate or defect are statistically significantly affected by credible threats of (intentional and random) sanctions (chisq(8)=38.6, p<0.01).

Table 4 provides estimation results for the three-alternative model where the statistically insignificant transfer and intention variables have been omitted, and Figure 4 describes the data that underlie that estimation. The figure shows, for each of the four cells, the fraction of subjects that cooperated, weakly non-cooperated and completely defected. (Recall that the number of observations, broken down by treatment, is provided in the legend to Figure 1.) When sanctions are not threatened, the most common trustee decision is to send some, but not all, of the amount that the investor requested. Such returns occur more than half of the time, and include 57% and 51% of all decisions in the low and high desired back transfer cases, respectively. Under sanction threats this weakly non-cooperative behavior becomes least common. About a fourth of trustees weakly non-cooperate under weak threats, but only 8% choose this option under severe threats.

The change in the frequency of non-cooperation underlies the impressive differences between the "threatened" and "not threatened" distributions described by Figure 4. The distribution in both of the "not threatened" cells has an inverted "U" shape, while in both "threatened" cells the shape of the distribution is the inverse. In fact, there is not a statistically significant difference between the two distributions in the "not threatened" cells. However, there is a statistically significant difference between any two other distributions in the remaining five pairwise comparisons (p<0.01 in all cases.) In particular, the change in shape from "U" to its inverse is statistically significant.

Although both weak and severe sanctions reduce weak non-cooperation, the effect of that reduction is different between the two cases. Threats of severe sanctions tend to

increase the number of cooperative decisions by trustees. When the desired back transfer is low and sanctions are not threatened we find that about 23% of the trustees return the amount that the investor requested. Under threats of severe sanctions the fraction of trustees who cooperate triples to 69%, and this difference is statistically significant (two sided, two-sample t-test with unequal variances, p=0.0).[10] Note that this effect is almost exclusively due to changes in the frequency of weak non-cooperation: the frequency of defection is nearly unchanged between the two treatments, at 19% and 22% in the no threat and threat cases, respectively.

When the threatened sanction is weak, the behavioral effect works in the opposite direction with the fraction of trustees who return nothing to their investor nearly doubling, from 25% without sanction threats to 46% under threats of weak sanctions. This increase is statistically significant (p=0.02) and is again primarily due to changes in the frequency of weak non-cooperation. In particular, the fraction of trustees who choose to cooperate is not statistically significantly different between the threat and no threat cases, at 30% and 25%, respectively.

It is worth emphasizing again that these results do not vary with investor intentions. The fraction of trustees who return nothing when threatened with weak sanctions is about 43% and 50% in the intentions and random conditions, respectively, and the difference is not statistically significant (p=0.58). Also, 72% of trustees cooperate when intentionally threatened with severe sanctions, while 65% do so when the severe sanction is imposed randomly by nature, and this difference is not significant (p=0.59).

IV.3.b. Five-alternative model

To further examine this effect on weak non-cooperation, we estimated a multinomial logit on a refined action space that includes five alternatives. There are three non-cooperation options: completely defect (send nothing), partially defect (return a positive amount less than the transfer amount), reciprocate (return at least as much as the transfer amount but less than the requested amount); and two cooperative options: cooperate

[10] Unless otherwise noted, all p-values reported below correspond to two-sided, two-sample t-tests, assuming unequal variances, of the null hypothesis that the means are the same.

(return exactly the requested amount) and strong cooperate (return more than the requested amount).[11]

Table 5 details the estimation results, and the underlying data are plotted in Figure 5. Note first that the frequency of strong cooperation is not statistically significantly affected by the threat of sanctions in either the high or low desired back transfer conditions. In the low case the frequencies are 12.2 and 14.9 with and without threats, respectively, while those frequencies are 5.3 and 4.6, respectively, for the high desired backtransfer case. Strong cooperation, as we have defined it, cannot be enforced by the sanctioning mechanism, and this might be a reason that its frequency does not vary with sanction threats. Also, we saw above that rates of complete defection with low desired backtransfer do not respond much to sanction threats, and because defection is defined the same way in this analysis the same is true here. Similarly, cooperation rates are not sensitive to weak sanction threats (p=0.55): 20.0% of trustees return exactly the requested amount when weak sanctions are not threatened, and 24.6% do so when they are.

Figure 4 indicated that threats of sanctions have a strong affect on the frequency of weak non-cooperation. Figure 5 refines this, and suggests that "reciprocity" (returning at least the investment amount but less than the requested amount) is substantially affected by threats in both the low and high desired back transfer condition. In the former case, the frequency of reciprocation is 17% and 0% in the no threat and threat conditions, respectively, and this difference is statistically significant (p=0.01). Similarly, in the weak punishment condition the change is from 27.7% to 5.3%, and again statistically significant (p<0.01).

Severe punishment also has an effect on the decision to return a positive amount that is less than the transferred amount: this is chosen by 40.4% of trustees in the low desired back transfer condition, but only 8.2% do so after being threatened with a severe sanction. This change is statistically significant (p<0.01.) There is not a significant effect

[11] It occurs a few times in our data that the desired back transfer is less than the transfer amount, meaning that a "cooperative" subject would in this case still return less than the transfer amount. In those few ambiguous cases we assigned the subject's decision to the most cooperative potential alternative. Our results are robust to dropping these observations, or to classifying them differently.

on this choice when the threat is weak: the frequency of this choice changes from 23% to 19% as a result of imposing threats of weak punishment (p=0.61).

Taken together, one interpretation of these findings is as follows. First, about 20% of trustees are dogmatic defectors who will send back nothing, and another 20% are dogmatic cooperators who will return the requested amount or more. These dogmatic decisions are insensitive to the presence of sanctions (or intentions).[12] The remaining 60% of trustees will make a return decision that is sensitive to whether sanctions have been threatened, and whether a threatened sanction is severe or weak. If punishment is not threatened, then roughly half of those 60% of trustees will return an amount at least as great as the investment amount, but less than the amount that the investor requested, and the remaining half will return a positive amount that is less than the investment amount. If, on the other hand, trustees are threatened with a severe sanction, then most of those sanction-sensitive 60% will choose to return exactly the requested amount. However, if that 60% are threatened with a weak sanction, then instead of allocating their return decisions roughly equally between reciprocation and partial defection, about half will instead choose to send back nothing, and most of the other half will choose to return a positive amount less than the investment amount.

IV.4. Comments on Unfair Backtransfer Request Data

There were 45 investors in our sample who requested an unfair backtransfer. As pointed out above, an "unfair" request is one for more than 2/3 of the tripled transfer amount. Table 1(B) shows that these requests are not balanced across cells: In the majority of these cases the unfair request was combined with a threat of a sanction. In the intention treatment we observed only four cases out of 27 where an unfair backtransfer request did not include the threat of punishment. It turns out that two-thirds (12 of 18) of the unfair requests in the incentive treatment were assigned to the sanction condition.

Although the sample size is small, there are nevertheless several features of this data worth noting. First, there were five cases where the investor chose to send one and

[12] The fact that the rates are insensitive does not necessarily imply that 40% of trustees are insensitive to our experiment's sanction incentives. However, Kurzban and Houser (2004, forthcoming), and Gunnthorsdottir et. al. (2001), among others, have found evidence from public goods experiments that subjects do differ systematically in their propensities to cooperate. At least in that environment, some subjects are dogmatic cooperators while others are dogmatic defectors.

ask for all three back. All of these cases occurred in the "intention" treatment, and in all cases the investors combined these decisions with a sanction threat. Because the sanction amount is four, the income-maximizing choice for the corresponding trustees is to send back the entire requested amount. Four of the five trustees did exactly this, and one sent back nothing. Table 1(B) provides statistics that include and do not include these five cases.

Second, after excluding the five "dominant strategy" cases just mentioned, one can compare means over the remaining 18 cases where punishment was intentional to the 12 cases where it was not. A casual inspection of the second and fifth columns of Table 1(B) reveals that behavior is not very different in those cases, indeed all pairwise tests reveal it to be identical. Again with the caveat that the sample size is small, we find that it is incentives, not intentions, which seem to be responsible for any sanction effect.

Consistent with Fehr and Rockenbach (2003), we find evidence that sanctions have a detrimental effect on cooperation. One can quickly see from Table 1(B), again excluding the dominant strategy cases and after pooling the intention and incentive treatments, that the amount returned when fines are imposed is about 14% and 12% of the desired back transfer and tripled transferred amounts, respectively. In the event that fines are not imposed these numbers are 29% and 25%, respectively.

V. Concluding Discussion

Threats of punishment are commonly used to encourage cooperation. Yet a substantial amount of cross-disciplinary research, and certainly casual observation, reveals that credible threats of punishment sometimes fail to foster cooperation. Both "intention" and incentive effects have been variously cited as reasons that punishment can fail. One goal of this research was to shed new light on the relative roles of incentives and intentions in reducing punishment's efficacy in a personal exchange environment. Moreover, because there are often a variety of ways in which one can behave non-cooperatively, our research also characterized differences in non-cooperative behavior among various incentive and intention conditions.

We reported data from human subjects who made decisions in a novel extension of an experiment reported by Fehr and Rockenbach (2003). We found that credible

threats often fail to produce cooperative behavior. This failure was not the result of intention effects: the effects of threats on trustworthiness did not depend on whether the threats were made intentionally by an investor or imposed randomly by nature. In particular, we found that trustees were statistically significantly more likely to return nothing to their investors after being threatened with weak punishment, regardless of whether the threat was made by the investor. This is consistent with the view that extrinsic incentives, regardless of how they are imposed, can crowd out internal incentives and, consequently, change subjects' cognitive frame from ethical to income maximizing (for related findings see, e.g. Kreps, 1997, Frey and Oberholzer-Gee, 1997; Tenbrunsel and Messick, 1999; Gneezy and Rustichini, 2000a,b).[13]

Trustees' decisions were substantially affected by the severity of the imposed punishment. We found that whether to choose severe punishment presents something of a dilemma for investors: using it provided a higher expected return but with increased variance. Weak punishment, on the other hand, somewhat reduced trustees' average return amounts, and statistically significantly increased the variance of the return distribution. Accordingly, threats of weak punishment are risky. A useful implication of this result is that, when designing sanctioning systems, policy makers need to consider not only the expected effect on cooperation, but also take into account the way non-cooperative behavior might be changed in the face of the new incentive structure, perhaps especially in the event that the threatened sanctions are weak.

Everyday examples of weak sanctioning systems are plentiful, and include fines for parking too long or not removing snow from sidewalks quickly enough. This might seem to call into question the external validity of our results. After all, if threats of weak sanctions are risky, and if people are generally risk averse, then one might expect evolutionary selection effects to diminish the use of weak sanctions over time. In fact, there are many reasons one might expect to see stable weak sanctioning systems in the naturally occurring world. One is that severe punishment is often not credible, because

[13] Convergent evidence of this effect is provided by ex post questionnaires that we administered to subjects after the experiment had concluded. These questionnaires, which were voluntary and not saliently rewarded, asked subjects why they made the decisions they did. Our casual examination of their responses suggested that it was common for trustees who were not threatened with sanctions to report that "guilt avoidance" was an important reason to return money to the investor. This sentiment was far less frequent in those cases where sanctions had been imposed. Regardless of whether the threat was intentional, threatened trustees were much more likely to say they made their decisions based on earnings considerations.

people expect officials will be unwilling to enforce extremely severe sanctions for minor violations. Moreover, even if a strong sanction is credible, it can be very expensive to enforce due to costs incurred in, for example, building and defending a case. Weak sanctions create the politically expedient impression that "something is being done" while at the same time having the great practical benefits of being both credible and relatively low cost to administer. And the difficulty in conducting counterfactual analyses can make the risks of weak sanction systems very difficult to quantify, or even detect.

Of course, one can point to examples of weak sanctioning systems that have been replaced. A prominent example of a large, complex and now abandoned weak sanctioning system is the Federal Communication Commissions' (FCC's) procedures to enforce indecency regulations. In principle, the FCC has always had the power to warn, sanction, suspend or even revoke the license of broadcasters who violate decency statutes. As a practical matter though, the FCC has never in its 75 year history suspended or revoked any station's broadcast license. Although the FCC did occasionally impose sanctions, they were infrequent and relatively low cost. For example, in 2000 the FCC levied exactly seven sanctions, for a total of $48,000, in response to complaints regarding indecent programming. Applied to this environment, our findings would suggest that broadcasters who found it in their financial best interest to violate the sanctions would systematically and repeatedly do so. Perhaps in part for this reason, the FCC has recently moved to replace the weak sanctioning system with one that makes credible threats of very severe sanctions. During the first seven months of 2004 the FCC levied over $1.5 million in new fines (Solomon, 2004).

This research focused on an environment where subjects made anonymous decisions exactly once. Consequently, our results cannot inform the way in which an ex-ante commitment to punish impacts the evolution of cooperation in repeat play. Does repeat play obscure the distinction between commitment and opportunity where punishers elect to offer cooperation on the next trial after punishing defection? Or does the ex ante commitment to punish spoil the reinforcement of cooperation in repeat interaction by creating a more "hostile" environment? In addition to a laboratory analysis geared towards those questions, future research might be profitably directed towards, for example, investigating satellite or internet radio broadcasters' programming decisions.

Satellite programming competes directly with FCC regulated broadcasts, but is itself neither licensed by nor subject to the regulations of the FCC. This could provide ideal complementary natural experiments on sanction effects.

References

Andreoni, J., Harbaugh, W. Vesterlund, L., (2003), "The Carrot or Stick: Reward, Punishment and Cooperation", *American Economic Review*, vol. 93, 893-902

Benabou, R., and Tirole, J., (2003), "Intrinsic and Extrinsic Motivation", Review of Economics Studies, 70, 489-520

Berg J., Dickhaut J., McCabe K. (1995), "Trust, Reciprocity and Social History", *Games and Economic Behavior*, 10, 122-142.

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

Blount, S. (1995): "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences", *Organizational Behavior and Human Decision Process* 63, 131-144.

Bolton, G., Brandts, J., Ockenfels, A., (1998). "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game." *Experimental Economics*, 1**,** 207-219.

Camerer, C. (2003). Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press.

Charness, G., Haruvy, E., (2002), "Altruism, Fairness, and Reciprocity in a Gift-Exchange Experiment: An Encompassing Approach", Games and Economic Behavior, 40, 203-231

Charness, Gary and Levine,David, " The Road to Hell: An Experimental Study of Intentions" (working paper)

Charness, G., and Dufwenberg, M., (2004) "Promises & Partnership" (working paper)

Deci, E.L., Koestner, R.M., Ryan, R., (1999), "A meta-analytic review of experiments examining the effect of extrinsic rewards on intrinsic motivation". Psychological Bulletin 125, 627–668.

Dufwenberg, M., Kirchsteiger, G., (2004), A theory of sequential reciprocity. Games & Economic Behavior 47, 268-98.

Dickinson,David L., (2001), "The Carrot vs. the Stick in Work Team Motivation", Experimental Economics, 4, 107-124

Falk, A., Fehr, E., Fischbacher, U., (2000): "Testing Theories of Fairness—Intentions Matter", Working Paper, University of Zurich

Falk, A., Gächter, S.,(1999), "Intrinsic Motivation and Extrinsic Incentives in a Repeated Game with Incomplete Contracts", Journal of Economic Psychology,20(3), 251-284

Falkinger, J., Fehr, E., Gächter,S., Winter-Ebmer, R, (2000): "A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence", *American Economic Review* ,90, 247-264

Fehr, E., (2000). "Cooperation and Punishment in Public Goods Experiments", American *Economic Review*, 90, 4, 980-994.

Fehr, E., Rockenbach, B. March, (2003). "Detrimental Effects of Sanctions on Human Altruism", *Nature* 422 ,137-140

Fehr, E., Gächter,S., (2002). "Do Incentive Contracts Undermine Voluntary Cooperation?". Working Paper, University of Zurich

Fehr, E., Klein, A. and Schmidt, K. M., "Contracts, Fairness, and Incentives", (working paper).

Fehr, E., Gächter, S., (2000): "Fairness and Retaliation: The Economics of Reciprocity", *Journal of Economic Perspectives*, 14(3), 159-181.

Fehr, E., Gächter, S., Kirchsteiger, G., (1997), "Reciprocity as a Contract Enforcement Device: Experimental Evidence", *Econometrica*, vol 65 (4), 833-860

Fehr, E., List, J., (2004) "The Hidden Costs and Rewards of Incentives", Journal of the European Economic Association, 2(5), 743-771.

Fehr, E. and Falk, A., (2002): "Psychological Foundation of Incentives", European Economic Review, 46, 687 - 724.

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

Gibbs, R. (1999), Intentions In The Experience Of Meaning, 69-106, Cambridge University Press.

Gneezy, U., Rustichini, A., (2000a), " A Fine Is A Price", *Journal of Legal Studies*, 29(1) , 1-17

Gneezy, U. and Rustichini, A., (2000b). Pay Enough or Don't Pay at All. *Quarterly Journal of Economics* 115(2), 791-810

Gordon, M., Bowlby. R. (1989). "Reactance and Intentionality Attribution as Determinants of the Intent to File a Grievance. *Personal Psychology*, 42, 309-329

Greenberg, M, and D. Frisch (1972). "Effect of Intentionality on Willingness to Return a Favor", *Journal of Experimental Social Psychology,* 8, 99-111

Houser, D., Keane, M. and McCabe, K. (2004), "Behavior in a Dynamic Decision Problem: An Analysis of Experimental Evidence Using a Bayesian Type Classification Algorithm", *Econometrica*. 72(3), 781-822.

Houser, D., Kurzban, R., (2002), "Revisiting Kindness and Confusion in Public Goods Experiments", *American Economic Review.* vol. 92,1062-1069.

Kreps, D., (1997), "Intrinsic Motivation and Extrinsic Incentives", American Economic Review, papers and Proceedings, 87(2), 359-364.

Kurzban, R., and Houser, D. (2004). "Experiments Investigating Cooperative Types in Humans: A Complement to Evolutionary Theory and Simulations", *Proceedings of the National Academy of Sciences of the United States of America*. Forthcoming.

Lepper, M. and Greene,D., (1978). The Hidden Cost of Reward: New Perspectives on the Psychology of Human Motivation, New York: John Wiley

McCabe, K., Rigdon, M. Smith, V., (2003): "Positive reciprocity and intention in trust games", *Journal of Economic Behavior and Organization*, Vol. 1523, 1-9

McCabe, K., Rassenti S. and Smith, V., (1996): "Game Theory and Reciprocity in Some Extensive Form Experimental Games", Proceedings of the National Academy of Sciences, 93, 13421-13428.

Nelson Jr., W.R., (2002). Equity and intention: it's the thought that counts*. Journal of Economic Behavior and Organization* , 48(4),423-430

Ostrom, Elinor, James Walker and Roy Gardner,1992, "Covenants With and Without a Sword: Self-Governance is Possible", *American Political Science Review* vol.86, 404 - 417.

Offerman, T., 2002: "Hurting Hurts More Than Helping Helps", *European Economic Review* 46, 1423-1437.

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

Schotter, A., Weiss, A.,Zapater, I., (1996): " Fairness and survival in ultimatum and dictatorship games". *Journal of Economic Behavior and Organization* 48 (4). 423-430

Sefton, M., Shupp,R. and Walker, J., (2002): "The Effects of Rewards and Sanctions in Provision of Public Goods by." (working paper).

Solomon, D., (2004): Transcript of prepared witness testimony before the House Committee on Energy and Commerce, Subcommittee on Telecommunications and the Internet. http://energycommerce.house.gov/108/Hearings/01282004hearing1165/Solomon1843.htm.

Taylor, C., (1979), Action as Expression, in Intention and Intentionality: Essays in Honour of G. E. M. Anscombe, Cora Diamond and Jenny Teichman, eds. Ithaca, N. Y.: Cornell University Press, 73-89.

Tenbrunsel, Ann E. and Messick, David M., (1999): "Sanctioning Systems, Decision Frames, and Cooperation", Administrative Science Quarterly, 44: 684-707

Jean-Robert Tyran and Lars P. Feld, (2004), "Achieving Compliance when Legal Sanctions are Non-Deterrent", working paper.

William Robert Nelson Jr.(2002): "Equity or intention: it is the thought that counts", Journal of Economic Behavior & Organization, vol.48, 423-430

Yamagishi, T., (1986): "The provision of a Sanctioning System as a Public Good." *Journal of Personality and Social Psychology*. Vol.51, No. 1, 110-116

Yamagishi, T., (1988): "Seriousness of Social Dilemmas and the Provision of a Sanctioning System*." Social Psychology Quarterly* Vol.51, No.1,32-42

Appendix A: The Model and Statistical Procedure

We model the trustee as making one choice among five mutually-exclusive alternatives. The three-alterative model also discussed in the body of the text can be developed analogously.[14] Of course, a trustee's action space includes more than five alternatives. Consequently, although both defensible and robust to reasonable changes, the mapping between subjects' actions and our model's alternatives is unavoidably arbitrary.

As described in the body of the text, the five alternatives we consider are (i) defect, (ii) partially defect, (iii) reciprocate, (iv) cooperate and (v) strongly cooperate. Each of the subjects' possible actions maps into a unique alternative in the model except in the empirically rare case where an investor requests a back-transfer that is not greater than the original transfer amount. This happens three times in our data when the investor sends zero (leaving the trustee with no choice to make), and two additional times when the transfer is non-zero. We have eliminated all of the cases where the trustee had no choice to make. In the two other cases the subject returned back exactly what the investor requested, and we have coded these subjects as cooperative. Our results do not change if we change their coding or drop them from the data set.

We model subjects' choices among these alternatives within a random utility model framework. An advantage of our procedure is that it is robust to the way in which we order alternatives. It is worthwhile to point out that it is inappropriate to use an ordered (e.g., ordered probit or ordered logit) framework to analyze these data. The reason is that there is no obvious way to order our model's alternatives. In particular, both the amount sent and amount requested determine the categorization of an individual's return decision, and it is not obvious how to map this two dimensional set to the real line. This is important because, in contrast with a random utility model, standard ordered frameworks include only one stochastic dimension. Consequently, there exist cross-alternative restrictions (on the way changes in one alternative's probability affect another's probability) that depend critically on the specific ordering chosen. While in many analyses it can be efficient to impose such restrictions, in our case it is not reasonable to report results about, for example, the effect of punishment on cooperation that rely heavily on the fact that defect is first and cooperation fourth on our list.

Our simple random utility model is developed as follows. We assume that the utility benefit for subject i in session s of each choice j can be modeled as a function of the nine-vector x_{ij} (whose elements are described in the body of the text) and a stochastic component ε_{isj} that depends on the individual, the session and the alternative. Letting β denote a real nine-vector, we then have:

$$U_{isj} = \beta' x_{ij} + \varepsilon_{isj} .$$

[14] Note that the three-alternative model is not a subset of the five-alternative model. The mapping between subjects' actions and the model's alternatives differs between the two cases.

Subject i is assumed to choose alternative j' if and only if $U_{isj'} > U_{isj}$ for all $j \neq j'$. We assume that ε_{isj} and $\varepsilon_{i's'j'}$ follow independent Weibull distributions whenever $s \neq s'$, and when $s = s'$ and $i = i'$. We allow for the possibility, however, that the error components are not independent among subjects within the same session.

These assumptions imply a standard multinomial logit framework, where the standard errors of our estimates are calculated using a White (1980) procedure that is robust to non-independence of the error-components within sessions.

Appendix B1: Instructions for the Intention Treatment

I. Investor's instruction

You are Actor 1

Description of Your Decision Problem

Thank you for coming! You've earned $5 for showing up on time, and the instructions explain how you can make decisions and earn more money. So please read these instructions carefully! There is no talking at any time during this experiment. If you have a question please raise your hand, and an experimenter will assist you.

You are in Room A and you will be randomly paired with someone in Room B. 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 your matched participant is in the role of Actor 2. You and Actor 2 will participate only once in this decision problem. You make your decisions with the help of the decision sheet described below.

This is how the experiment works.

Endowment

Besides the $5 show-up bonus, at the beginning of the experiment **both** actors receive an initial endowment of 10 **E**$ (experimental dollars).

Your decision

Your decision includes three parts:

1. A transfer between 0 and 10 E$ to Actor 2.

You, as Actor 1, can transfer, from your endowment, any amount between 0 and 10 E$ to Actor 2. The experimenters will **triple** this transferred amount, so that Actor 2 receives three times the amount of E$ you transferred.

2. A desired back-transfer.

You also make a decision about your desired back-transfer, that is, at least how many E$ you would like to receive back from Actor 2. You can ask for any amount between zero and the tripled amount of your transfer.

3. **Whether to impose a conditional payoff cut of 4 E$ on Actor 2's final earnings.**
    - A conditional payoff cut of 4 E$ for Actor 2 has the following consequences. The payoff of Actor 2 will be reduced by 4 E$ if his/her actual back-transfer is less than your desired back-transfer. The conditional payoff cut does not happen if Actor 2 transfers your desired amount or more to you.
    - If you choose not to impose a conditional payoff cut, then the income of Actor 2 will not be reduced, irrespective of the amount of Actor 2's back-transfer.

The decision of Actor 2

After your decision, Actor 2 can transfer back to you any amount of the tripled number of E$ bills he/she received. In case that you have chosen a conditional payoff cut of 4 E$, and if Actor 2 transfers back less than what you desired, then Actor 2 must pay the conditional cut.

Payoffs

You (Actor 1) receive: 10 E$ – transfer to Actor 2 + back-transfer from Actor 2.

Actor 2 receives**:** 10 E$ + 3 × transfer from Actor 1 – back-transfer to Actor 1 – 4 E$ (in case that a conditional payoff cut was imposed and is due)

Exchange rate: For every E$ you earn you will be paid $1.

How the experiment is conducted

There are several envelopes in Room A and Room B. In each envelope in Room A and Room B, there is a card marked with a unique letter. Each envelope looks the same. Everyone in both Room A and Room B will randomly pick up an envelope. The person in Room B who chooses the card with the same letter as yours will be your Actor 2.

Items on your table: 10 E$ bills (your endowment), two decision sheets (one for Actor 1 and one for Actor 2) and two **Yes/No** stickers.
Items on Actor 2's table: 10 E$ bills (Actor 2's endowment)

You will make your decisions at your seat by filling in the decision sheets. You need to leave the number of E$ bills you want to transfer in the envelope, but keep the card, which will help the experimenter to return the envelope to you later. Raise your hand after you're done. The experimenter will go to your seat, check whether all necessary information is on the decision sheets, then triple your transferred E$ bills. The experimenter will also record the letter of your card on the back of the decision sheets so that your envelope can be given to your Actor 2 who has the card with the same letter. The experimenter will then put the decision sheets and tripled amount of E$ bills in your envelope and collect it. After every Actor 1 has finished, the experimenter will take all the envelopes to Room B.

The experimenter will give each Actor 1's envelope to his/her Actor 2. Each Actor 2 then decides how much to transfer back to you by writing down a number on the decision sheets and leave the E$ bills he/she wants to transfer back to you in the envelope. When Actor 2 has finished his/her decision, the experimenter will go to his/her seat, check whether all the necessary information is on the decision sheets and then put your copy of the decision sheet in the envelope. Actor 2 will keep his/her copy of the decision sheet.

After all Actor 2s have finished, the experimenter will take all the envelopes to Room A. The experimenter will return to you the envelope with the back-transfer E$ bills from Actor 2 inside. Each Actor 1 will then be called, one by one, to the experimenter. When

called, you will take your envelope to the experimenter. The experimenter will calculate your final earnings and pay you privately. Then please exit the lab. Since you will be asked to leave when you are done, you should take all your belongings when you go to the experimenter. Actor 2s in room B will be paid after **all** Actor 1s have been paid and have left the lab.

Throughout this experiment, you won't meet any Actor 2 in room B.

End of Instructions

Please raise your hand to indicate that you are finished reading these instructions. When you do, an experimenter will give you a short quiz to ensure that you understand how you make decisions.

II. Trustee's instruction

You are Actor 2

Description of Your Decision Problem

Thank you for coming! You've earned $5 for showing up on time, and the instructions explain how you can make decisions and earn more money. So please read these instructions carefully! There is no talking at any time during this experiment. If you have a question please raise your hand, and an experimenter will assist you.

You are in Room B and you will be randomly paired with someone in Room A. 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 your matched participant is in the role of Actor 1. You and Actor 1 will participate only once in this decision problem. You make your decisions with the help of the decision sheet described below.

This is how the experiment works.

Endowment

Besides the $5 show up bonus, at the beginning of the experiment **both** actors receive an initial endowment of 10 **E**$ (experimental dollars).

The decision of Actor 1 (You are not Actor 1)

First Actor 1 has to make a decision that consists of the following three components.

4. A transfer between 0 and 10 E$ to you.

Actor 1 can transfer, from his/her endowment, any amount between 0 and 10 E$ to you. The experimenters will **triple** this transferred amount, so that you receive three times the amount of E$ transferred by Actor 1.

5. A desired back-transfer.

Actor 1 will indicate his/her desired back-transfer, which is at least how many E$ he/she would like to receive back from you. Actor 1 can ask for any amount between zero and the tripled amount of his/her transfer.

6. **Whether to impose a conditional payoff cut of 4 E$ on your final earnings.**
    - A conditional payoff cut of 4 E$ has the following consequences for you. Your payoff will be reduced by 4 E$ if your actual back-transfer is less than the back-transfer desired by Actor 1. The conditional payoff cut does not happen if you transfer the desired amount or more to Actor 1.
    - If Actor 1 chooses not to impose a conditional payoff cut, then your income will not be reduced, irrespective of the amount of your back-transfer to Actor 1.

Your decision

After Actor 1 makes his/her decisions, you, as Actor 2, can transfer back to Actor 1 any amount of the tripled number of E$ bills you received. As noted, in case that Actor 1 has chosen a conditional payoff cut of 4 E$, and you transfer back less than what he/she desired, then you must pay the conditional cut.

Payoffs

Actor 1 receives: 10 E$ – transfer to Actor 2 + back-transfer from Actor 2.

You (Actor 2) receive**:** 10 E$ + 3× transfer from Actor 1 – back-transfer to Actor 1 – 4 E$ (in case that a conditional payoff cut was imposed and is due)

Exchange rate: For every E$ you earn you will be paid $1.

How the experiment is conducted

There are several envelopes in Room A and Room B. In each envelope in Room A and Room B, there is a card marked with a unique letter. Each envelope looks the same. Everyone in both Room A and Room B will randomly pick up an envelope. The person in Room A who chooses the card with the same letter as yours will be your Actor 1.

Items on your table: 10 E$ bills (Your endowment).
Items on Actor 1's table: 10 E$ bills (Actor 1's endowment), two decision sheets (one for Actor 1 and one for Actor 2, as shown below) and two **Yes/No** stickers.

Sample Decision Sheets

Decision Sheet Copy for **Actor 1**

Actor 1:

1. I decide to transfer _____E$ to Actor 2
2. My desired back-transfer amount: _____E$
3. If Actor 2's back-transfer is less than my desired back-transfer amount, I will impose a conditional payoff cut of 4 E$ on Actor 2:

Yes No

Actor 2:

Based on Actor 1's decision, I decide to transfer _____ E$ back to Actor 1

Decision Sheet Copy for **Actor 2**

Actor 1:

1. I decide to transfer _____E$ to Actor 2
2. My desired back-transfer amount: _____E$
3. If Actor 2's back-transfer is less than my desired back-transfer amount, I will impose a conditional payoff cut of 4 E$ on Actor 2:

Yes No

Actor 2:

Based on Actor 1's decision, I decide to transfer _____ E$ back to Actor 1

Actor 1 will make his/her decision at his/her seat by filling in the decision sheets. Actor 1 will leave the number of E$ bills he/she wants to transfer in the envelope he/she picked up, but keep the card, which will help the experimenter to return the envelope to him/her later. When Actor 1 is done, the experimenter will go to his/her seat, check whether all necessary information is on the decision sheets, then triple the transferred E$ bills. The experimenter will also record the letter of his/her card on the back of the decision sheets so that his/her envelope can be given to his/her Actor 2 who has the card with the same letter. The experimenter will then put the decision sheets and tripled number of E$ bills in his/her envelope and collect it. After every Actor 1 has finished, the experimenter will take all the envelopes to Room B.

The experimenter will give each of you your Actor 1's envelope. When you get the envelope, decide how much to transfer back to Actor 1 by writing down a number on the decision sheets and leave the E$ bills you want to transfer back to Actor 1 in the envelope. Raise your hand when you are done. The experimenter will go to your seat, check whether all the necessary information is on the decision sheets and then put Actor 1's copy of the decision sheet in the envelope and collect it. You will keep your copy of the decision sheet. Don't show anybody else your decision sheet.

After all Actor 2s have finished, the experimenter will take all the envelopes to Room A. The experimenter will return to Actor 1 his/her envelope with the back-transfer E$ bills from Actor 2 inside. Each Actor 1 will be called, one by one, to the experimenter. The experimenter will calculate his/her final earnings and pay him/her privately. Then Actor 1 will exit the lab. After all Actor 1s have left, the experimenter will call each Actor 2 one by one. When called, you will go to the experimenter with your decision sheet. The experimenter will calculate your earnings and pay you privately.

Throughout this experiment, you won't meet any Actor 2 in room B.

End of Instructions

Please raise your hand to indicate that you are finished reading these instructions. When you're done, an experimenter will give you a short quiz to ensure that you understand how you make decisions.

Appendix B2. Instructions for the Random Treatment

I. Investor's instruction

You are Actor 1

Description of Your Decision Problem

Thank you for coming! You've earned $5 for showing up on time, and the instructions explain how you can make decisions and earn more money. So please read these instructions carefully! There is no talking at any time during this experiment. If you have a question please raise your hand, and an experimenter will assist you.

You are in Room A and you will be randomly paired with someone in Room B. 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 your matched participant is in the role of Actor 2. You and Actor 2 will participate only once in this decision problem. You make your decisions with the help of the decision sheet described below.

This is how the experiment works.

Endowment

Besides the $5 show-up bonus, at the beginning of the experiment **both** actors receive an initial endowment of 10 **E**$ (experimental dollars).

Your decision

Your decision includes two parts:

1. **A transfer between 0 and 10 E$ to Actor 2.**
   You, as Actor 1, can transfer, from your endowment, any amount between 0 and 10 E$ to Actor 2. The experimenters will **triple** this transferred amount, so that Actor 2 receives three times the amount of E$ you transferred.

2. **A desired back-transfer.**
   You also make a decision about your desired back-transfer, that is, at least how many E$ you would like to receive back from Actor 2. You can ask for any amount between zero and the tripled amount of your transfer.

The decision of Actor 2

After your decision, Actor 2 can transfer back to you any amount of the tripled number of E$ bills he/she received.

Payoffs

You (Actor 1) receive: 10 E$ – transfer to Actor 2 + back-transfer from Actor 2.

Exchange rate: For every E$ you earn you will be paid $1.

How the experiment is conducted

There are several envelopes in Room A and Room B. In each envelope in Room A and Room B, there is a tag marked with a unique letter. Each envelope looks the same. Everyone in both Room A and Room B will randomly pick up an envelope. The person in Room B who chooses the tag with the same letter as yours will be your Actor 2.

Items on your table: 10 E$ bills (your endowment), two decision sheets (one for Actor 1 and one for Actor 2).
Items on Actor 2's table: 10 E$ bills (Actor 2's endowment)

You will make your decisions at your seat by filling in the decision sheets. You need to leave the number of E$ bills you want to transfer in the envelope, but keep the tag, which will help the experimenter to return the envelope to you later. Raise your hand after you're done. The experimenter will go to your seat, check whether all necessary information is on the decision sheets, then triple your transferred E$ bills. The experimenter will also record the letter of your tag on the back of the decision sheets so that your envelope can be given to your Actor 2 who has the tag with the same letter. The experimenter will then put the decision sheets and tripled amount of E$ bills in your envelope and collect it. After every Actor 1 has finished, the experimenter will take all the envelopes to Room B.

Each Actor 2 will get his/her Actor 1's envelop according to the letter on their tags. Each Actor 2 then decides how much to transfer back to you by writing down a number on the decision sheets and leave the E$ bills he/she wants to transfer back to you in the envelope. When Actor 2 has finished his/her decision, the experimenter will go to his/her seat, check whether all the necessary information is on the decision sheets and then put your copy of the decision sheet in the envelope. Actor 2 will keep his/her copy of the decision sheet.

After all Actor 2s have finished, the experimenter will take all the envelopes to Room A. The experimenter will return to you the envelope with the back-transfer E$ bills from Actor 2 inside. Each Actor 1 will then be called, one by one, to the experimenter. When called, you will take your envelope to the experimenter. The experimenter will calculate your final earnings and pay you privately. Then please exit the lab. Since you will be asked to leave when you are done, you should take all your belongings when you go to the experimenter. Actor 2s in room B will be paid after **all** Actor 1s have been paid and have left the lab.

Throughout this experiment, you won't meet any Actor 2 in room B.

End of Instructions

Please raise your hand to indicate that you are finished reading these instructions. When you do, an experimenter will give you a short quiz to ensure that you understand how you make decisions.

I. Trustee's instruction

You are Actor 2

Description of Your Decision Problem

Thank you for coming! You've earned $5 for showing up on time, and the instructions explain how you can make decisions and earn more money. So please read these instructions carefully! There is no talking at any time during this experiment. If you have a question please raise your hand, and an experimenter will assist you.

You are in Room B and you will be randomly paired with someone in Room A. 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 your matched participant is in the role of Actor 1. You and Actor 1 will participate only once in this decision problem. You make your decisions with the help of the decision sheet described below.

This is how the experiment works.

Endowment

Besides the $5 show up bonus, at the beginning of the experiment **both** actors receive an initial endowment of 10 **E**$ (experimental dollars).

The decision of Actor 1 (You are not Actor 1)

First Actor 1 has to make a decision that consists of the following two components.

7. A transfer between 0 and 10 E$ to you.

Actor 1 can transfer, from his/her endowment, any amount between 0 and 10 E$ to you. The experimenters will **triple** this transferred amount, so that you receive three times the amount of E$ transferred by Actor 1.

8. A desired back-transfer.

Actor 1 will indicate his/her desired back-transfer, which is at least how many E$ he/she would like to receive back from you. Actor 1 can ask for any amount between zero and the tripled amount of his/her transfer.

The randomly determined Conditional Payoff-Cut

Half of the Actor 2s will be randomly assigned to receive the Payoff-Cut and **half** randomly assigned not to receive the Payoff-Cut.

- If you are randomly assigned to the Payoff-Cut, there will be a conditional payoff cut of 4 E$ for you. A conditional payoff cut has the following consequences: Your payoff will be reduced by 4 E$ if your actual back-transfer is less than Actor 1's desired back-transfer. The conditional payoff cut does not happen if you transfer the desired amount or more to Actor 1.

- If you are randomly assigned to No Payoff-Cut, then your earnings will not be reduced, irrespective of the amount of your back-transfer.

Important:

(1) Actor 1s have not been told about the conditional payoff cut. When Actor 1s make their transfer and desired back-transfer decisions, they don't know that some Actor 2s will be assigned to a conditional payoff-cut.

(2) Whether you are assigned to the Payoff-Cut is randomly determined. Nothing that you or Actor 1 does affects whether you are assigned to the Payoff Cut. Each Actor 2's assignment is indicated by the Payoff-Cut sticker and No Payoff-Cut sticker (as explained below).

(3) At the end of the experiment, *after all decisions have been made*, Actor 1s will be informed about the Payoff-Cut, and will be told whether their matched Actor 2s were randomly assigned to the Payoff-Cut.

Your decision

After Actor 1 makes his/her decisions, and after whether you are assigned to the conditional payoff-cut has been randomly determined, you, as Actor 2, can transfer back to Actor 1 any amount of the tripled number of E$ bills you received. As noted, in case that you are assigned to the conditional payoff cut of 4 E$, and you transfer back less than what he/she desired, then you must pay the conditional cut.

Payoffs

Actor 1 receives: 10 E$ – transfer to Actor 2 + back-transfer from Actor 2.

You (Actor 2) receive**:** 10 E$ + 3× transfer from Actor 1 – back-transfer to Actor 1 – 4 E$ (in case that a conditional payoff-cut is due).

Exchange rate: For every E$ you earn you will be paid $1.

How the experiment is conducted

There are several envelopes in Room A and Room B. In each envelope in Room A and Room B, there is a tag marked with a unique letter. Each envelope looks the same. Everyone in both Room A and Room B will randomly pick up an envelope. The person in Room A who chooses the tag with the same letter as yours will be your Actor 1.

Items on your table: 10 E$ bills (Your endowment).
Items on Actor 1's table: 10 E$ bills (Actor 1's endowment), two decision sheets (one for Actor 1 and one for Actor 2, as shown below).

Sample Decision Sheets

Decision Sheet Copy for **Actor 1**

Actor 1:
1. I decide to transfer _____E$ to Actor 2
2. My desired back-transfer amount: _____E$

Actor 2:
I decide to transfer _____ E$ back to Actor 1

Decision Sheet Copy for **Actor 2**

Actor 1:
1. I decide to transfer _____E$ to Actor 2
2. My desired back-transfer amount: _____E$

Actor 2:
I decide to transfer _____ E$ back to Actor 1

Actor 1 will make his/her decision at his/her seat by filling in both of the decision sheets. Actor 1 will leave the number of E$ bills he/she wants to transfer in the envelope he/she picked up, but keep the tag, which will help the experimenter to return the envelope to him/her later. When Actor 1 is done, the experimenter will go to his/her seat, check whether all necessary information is on the decision sheets, then triple the transferred E$ bills. The experimenter will also record the letter of his/her tag on the back of the decision sheets so that his/her envelope can be given to his/her Actor 2 who has the tag with the same letter. The experimenter will then put the decision sheets and tripled number of E$ bills in his/her envelope and collect it. After every Actor 1 has finished, the experimenter will take all the envelopes to Room B.

In Room B, the experimenter will first randomly choose half of the envelopes. For each envelope, the experimenter will take out both Decision sheets, and put a Payoff-Cut sticker (as shown below) on both decision sheets and put it back into the envelope. Similarly, for the other half of the envelopes, the experimenter will put No Payoff-Cut sticker on both decision sheets.

Payoff-Cut sticker

Randomly Determined Payoff-Cut

If Actor 2's back-transfer is less than Actor 1's desired back-transfer amount, there will be a conditional payoff cut of 4 E$ on Actor 2:

Yes

No Payoff-Cut sticker

Randomly Determined Payoff-Cut

If Actor 2's back-transfer is less than Actor 1's desired back-transfer amount, there will be a conditional payoff cut of 4 E$ on Actor 2:

No

After the conditional payoff-cut has been randomly assigned, the experimenter will give each Actor 2 his/her Actor 1's envelope. When you get the envelope, the sticker on the decision sheets will tell you whether you have been randomly assigned to the conditional payoff-cut. Based on this and Actor 1's decision, you will decide how much to transfer back to Actor 1. You will write the amount you want to transfer on both decision sheets, and also place the E$ bills you want to transfer to Actor 1 in the envelope.

Raise your hand when you are done. The experimenter will go to your seat, check whether all the necessary information is on the decision sheets and then put Actor 1's copy of the decision sheet in the envelope and collect it. You will keep your copy of the decision sheet. Don't show anybody else your decision sheet.

After all Actor 2s have finished, the experimenter will take all the envelopes to Room A. The experimenter will explain the Conditional Payoff Cut to the Actor 1s at this time. *This is the first time that Actor 1s will learn about the Payoff Cut.* The experimenter will return to Actor 1 his/her envelope with the decision sheet and the back-transfer E$ bills from Actor 2 inside. Each Actor 1 will be called, one by one, to the experimenter. The experimenter will calculate his/her final earnings and pay him/her privately. Then Actor 1 will exit the lab. After all Actor 1s have left, the experimenter will call Actor 2s one by one. When called, you will go to the experimenter with your decision sheet. The experimenter will calculate your earnings and pay you privately.

Throughout this experiment, you won't meet any Actor 1 in room A.

End of Instructions

Please raise your hand to indicate that you are finished reading these instructions. When you're done, an experimenter will give you a short quiz to ensure that you understand how you make decisions.

Table 1. (A) Mean decisions by investors and trustees when request is fair*

	Random Treatment		Intention Treatment	
	No threats imposed	Threats imposed	No threats imposed	Threats imposed
Investment	6.2	5.5	6.2	5.3
	(0.4)	(0.5)	(0.3)	(0.4)
Desired back-transfer as a percentage of tripled investment	57.0	60.2	55.3	59.0
	(1.7)	(1.3)	(1.5)	(1.2)
Actual back-transfer	5.7	4.3	5.2	4.2
	(0.8)	(0.8)	(0.7)	(0.6)
Actual back-transfer as a percentage of tripled investment	29.1	31.8	27.2	31.9
	(3.2)	(4.9)	(3.0)	(3.8)
Actual back-transfer as a percentage of request	53.4	53.1	54.5	56.3
	(6.0)	(7.9)	(8.0)	(6.7)
Number of observations	52 pairs	44 pairs	60 pairs	62 pairs

(B) Mean decisions by investors and trustees when request is unfair

	Random Treatment		Intention Treatment		
	No threats imposed	Threats imposed	No threats imposed	Threats imposed	
				All	Exclude (1,3)**
Investment	5.3	5.8	7.0	5.2	6.3
	(1.3)	(0.8)	(1.1)	(0.7)	(0.6)
Desired back-transfer as a percentage of tripled investment	91.2	83.2	91.9	92.3	90.1
	(5.6)	(3.2)	(6.7)	(2.0)	(2.3)
Actual back-transfer	5.3	2.5	5.0	2.2	2.2
	(2.5)	(1.4)	(1.5)	(0.8)	(0.9)
Actual back-transfer as a percentage of tripled investment	23.4	12.1	27.0	26.5	11.6
	(10.0)	(6.7)	(8.0)	(8.1)	(4.7)
Actual back-transfer as a percentage of request	29.0	14.4	30.7	28.3	14.0
	(13.3)	(8.4)	(9.5)	(8.3)	(5.7)
Number of observations	6 pairs	12 pairs	4 pairs	23 pairs	18 pairs

*Numbers in parentheses are standard errors.

**Here investors transfer one, request three back and impose the fine of four. There are five observations in this case. Of these, one trustee returned zero and four returned three.

Table 2: Effect of Intentions and Incentives on Percentage of Tripled Investment Amount Returned by Trustees

Variables	Coefficient
High investment(=1 investment≥4;=0, o.w)	-7.55
	(5.25)
Intention threat	28.51
	(6.83)
Random threat	31.06
	(8.10)
Intention no threat	35.28
	(6.56)
Random no threat	39.22
	(6.24)
Random threat and low request	13.39
	(11.11)
Intention threat and low request	18.52
	(7.19)
Intention no threat and low request	-4.22
	(7.02)
Random no threat and low request	-9.64
	(6.47)
R^2	0.5913

Numbers in parentheses are standard errors;
Low request: desired back transfer<8 and Higher request: desired backtransfer ≥ 8

Table 3: Effect of Intentions and Incentives on Non-Cooperation (Three-alternative Model)

Variables	Complete Defect (Return=0)	Complete Cooperate (Return≥Request)
	Coefficient	Coefficient
High investment(=1 if investment≥4;=0, o.w)	-0.43	-0.26
	(0.58)	(0.64)
Intentional threat	0.67	-0.06
	(0.63)	(0.86)
Random threat	1.82	1.36
	(0.82)	(1.04)
Intentional no threat	-0.16	-0.69
	(0.65)	(0.77)
Random no threat	-0.49	-0.25
	(0.66)	(0.75)
Random threat and low request	0.13	1.31
	(1.30)	(1.35)
Intentional threat and low request	-0.01	2.10
	(0.85)	(1.03)
Intentional no threat and low request	-0.55	0.22
	(0.63)	(0.82)
Random no threat and low request	-0.54	-0.86
	(0.64)	(0.82)
Pseudo-R^2	0.1280	

Numbers in parentheses are standard errors.
Baseline: Weakly non-cooperate (0<Return<Request)

Table 4: Effect of Punishment Threat on Non-Cooperation (Three-alternative Model)

Variables	Complete Defect (Return=0)	Complete Cooperate (Return≥Request)
	Coefficient	Coefficient
Threat	0.62	0.19
	(0.25)	(0.37)
No threat	-0.72	-0.72
	(0.25)	(0.30)
Threat and low request	0.39	1.95
	(0.62)	(0.66)
No threat and low request	-0.37	-0.17
	(0.42)	(0.49)
Pseudo-R^2	0.1125	

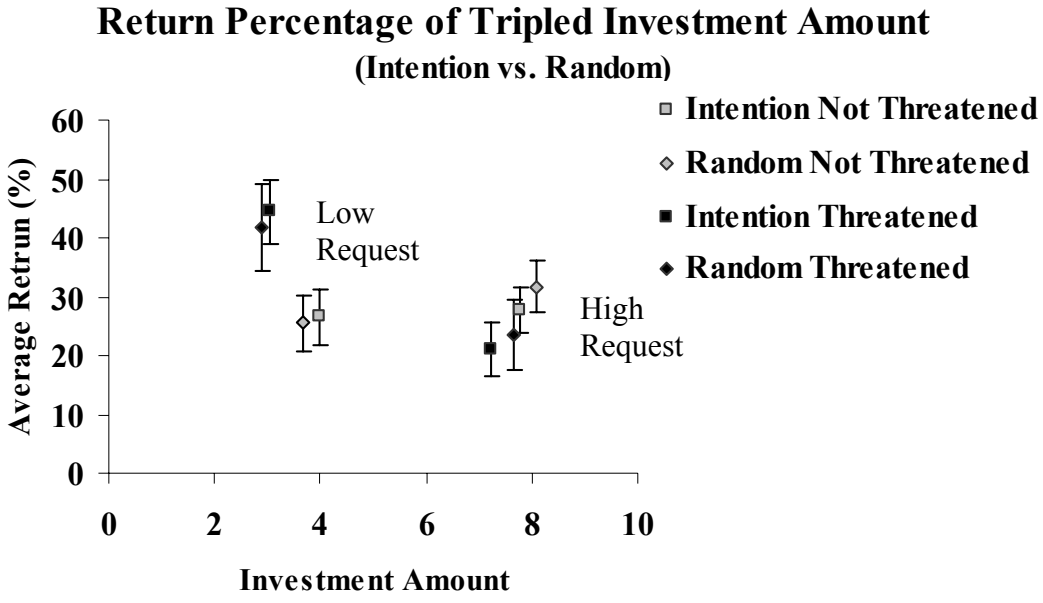
Numbers in parentheses are standard errors.
Baseline: Weakly non-cooperate (0<Return<Request)

Table 5: Effect of Punishment Threats on Non-Cooperation
(Five-alternative Model)

Variables	Partially defect (0<Return<Investment)	Reciprocate (Investment≤Return<Request)	Cooperate (Return=Request)	Strong Cooperate (Return>Request)
	Coefficient	Coefficient	Coefficient	Coefficient
Threat	-0.86	-2.16	-0.62	-2.16
	(0.28)	(0.52)	(0.29)	(0.57)
No threat	-0.06	0.12	-0.21	-1.67
	(0.34)	(0.26)	(0.38)	(0.65)
Threat and low request	-0.15	-31.67	1.55	1.55
	(0.61)	(0.65)	(0.46)	(0.82)
No threat and low request	0.81	-0.24	-0.60	1.42
	(0.56)	(0.43)	(0.65)	(0.92)
Pseudo-R^2		0.1516		

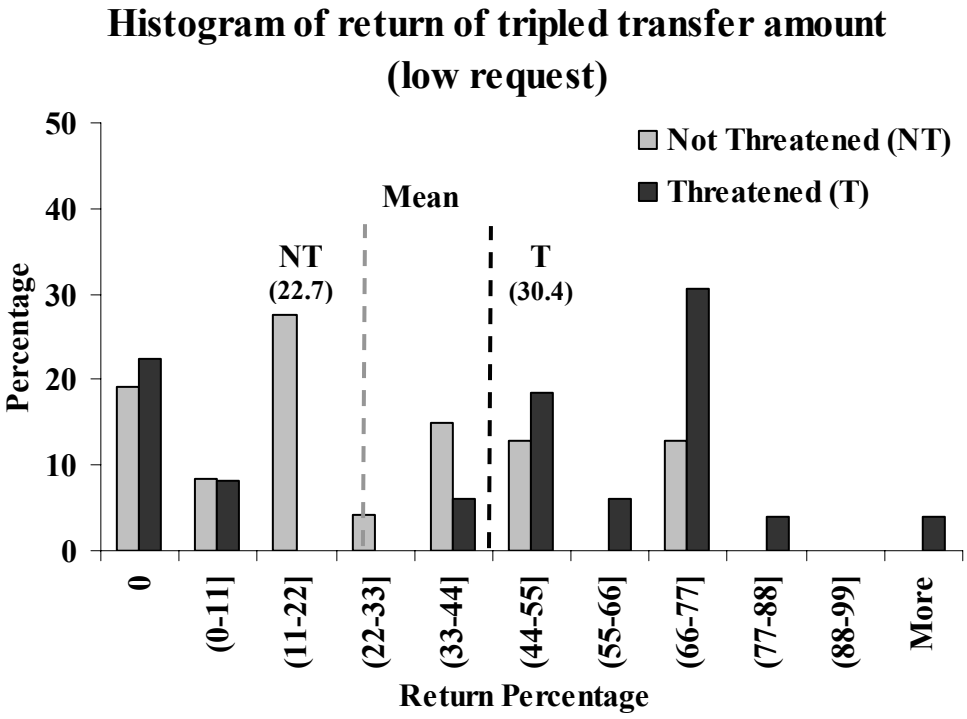
Numbers in parentheses are standard errors. Baseline: Completely defect (Return=0)

Figure 1



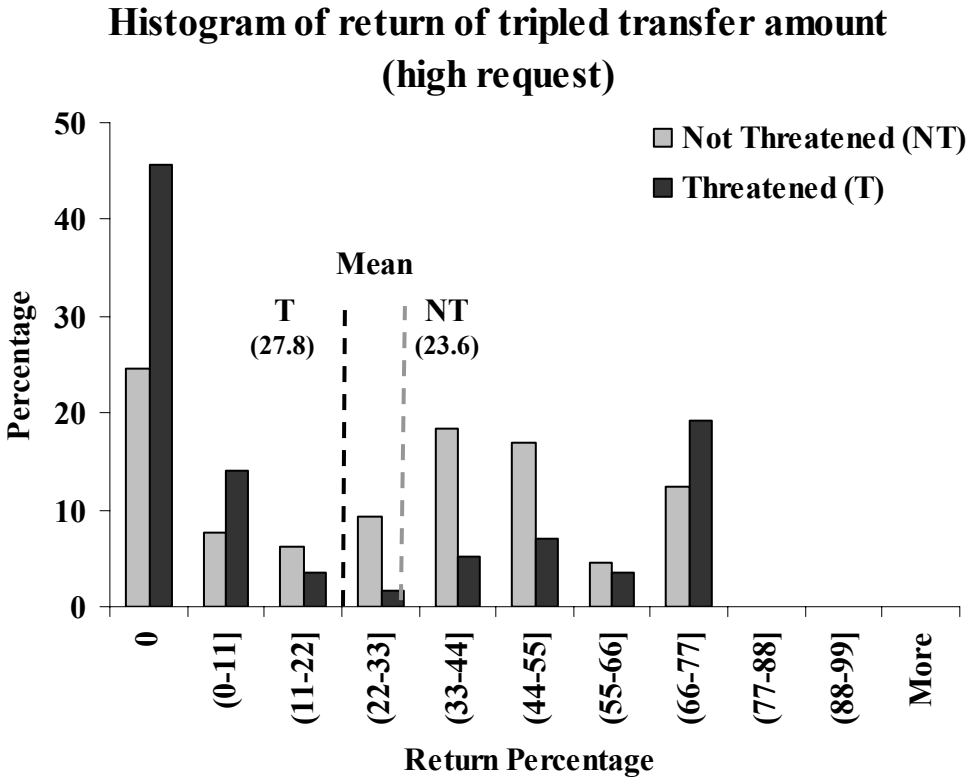
Let L denote "Low Request," H denote "High Request," N="Not Threatened," T="Threatened," R="Random" and I="Intention." Then the number of trustee observations in each cell is as follows. LRN=22, LIN=25, LRT=20, LIT=29, HRN=30, HIN=35, HRT=24, HIT=33.

Figure 2



Numbers in parentheses are standard deviations of the respective distributions.

Figure 3



Numbers in parentheses are standard deviations of the respective distributions.

Figure 4

Distribution of Trustees' Decisions (Three Alternatives)

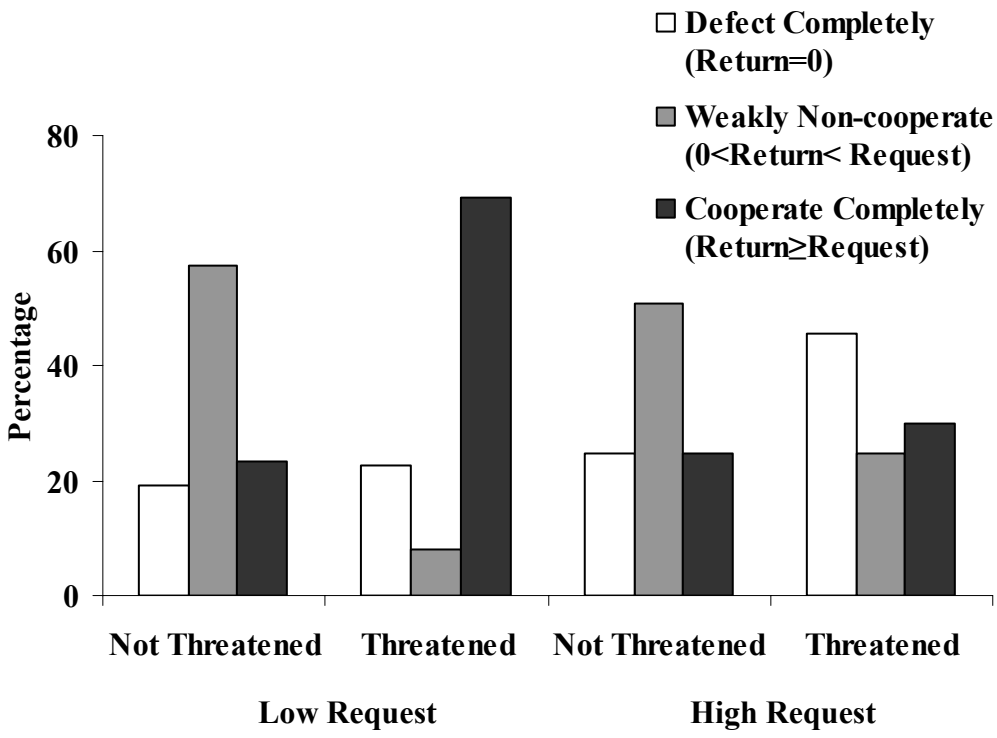
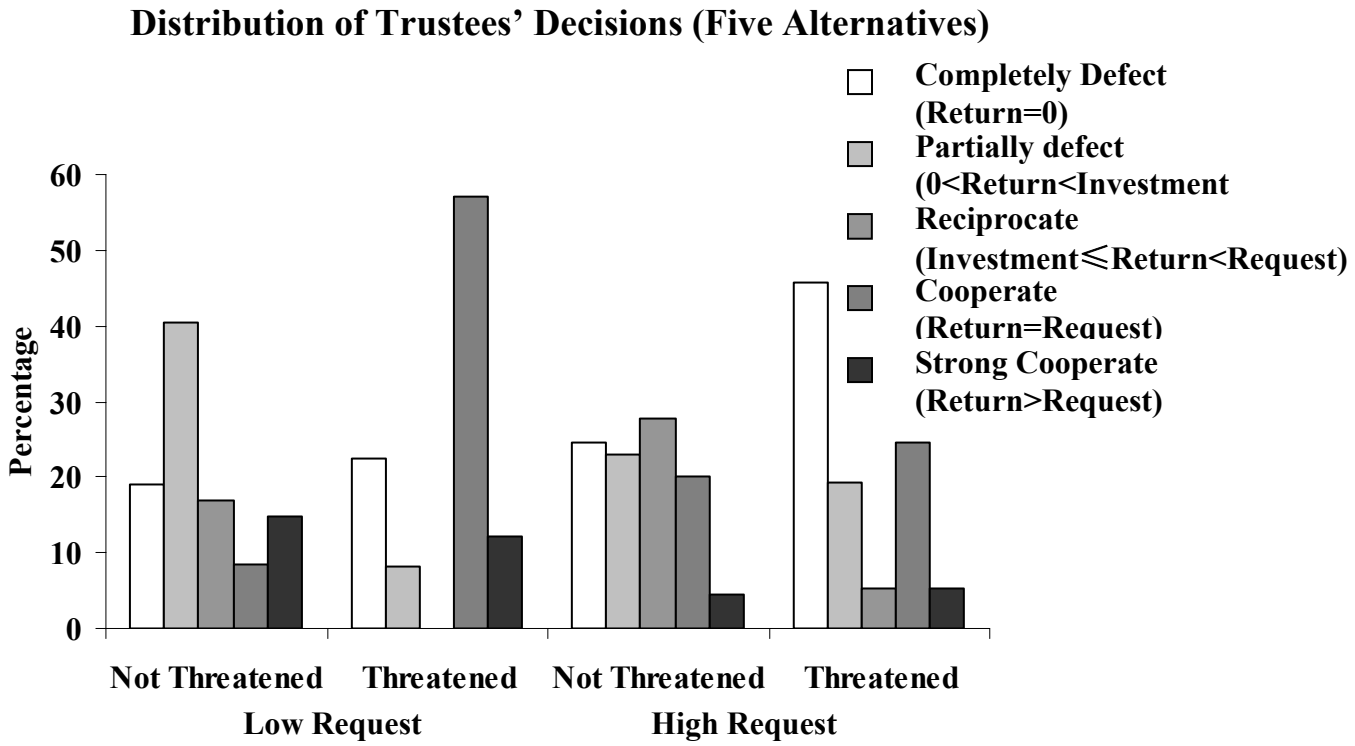


Figure 5


Distribution of Trustees' Decisions (Five Alternatives)
Completely Defect (Return=0)
Partially defect (0<Return<Investment
Reciprocate (Investment≤Return<Request)
Cooperate (Return=Request)
Strong Cooperate (Return>Request)
Percentage
60
50
40
30
20
10
0
Not Threatened
Threatened
Not Threatened
Threatened
Low Request
High Request