TRUTH OR CONSEQUENCES: AN EXPERIMENT*

Jordi Brandts and Gary Charness+

October, 2002

Abstract: This paper presents evidence that the willingness to punish an unfair action is sensitive to whether this action was preceded by a deceptive message. One player first sends a message indicating an intended play, which is either favorable or unfavorable to the other player in the game. After the message, the sender and the receiver play a simultaneous 2x2 game, in which the sender may or may not play according to his message. Outcome cells may, hence, be reached following true or false messages. In the third stage the receiver may (at a cost) punish or reward, depending on which cell of the simultaneous game has been reached. We test whether receivers' rates of monetary sacrifice depend on the process by which an outcome is reached. We study two decision-elicitation methods: the strategy and the direct response methods. For each method, deception more than doubles the punishment rate as a response to an action that is unfavorable to the receiver. We also find evidence that 17-25% of all participants choose to reward a favorable action choice made by the sender, even though doing so leaves one at a payoff disadvantage. Our results reflect on current economic models of utility and have implications for organizational decision-making behavior.

* We thank Antonio Cabrales, Rachel Croson, Maurice Schweitzer, seminar participants at the University of Amsterdam and the University of Valencia, and three anonymous reviewers for helpful comments. Brandts gratefully acknowledges support from the Spanish DGCICYT (PB98-0465). This project was started while he was visiting the Department of Economics of the University of California at Berkeley. He thanks the members of the Department at Berkeley for their hospitality. Charness gratefully acknowledges support from the MacArthur Foundation. This project was started while he was affiliated with Universitat Pompeu Fabra in Barcelona. This paper is part of the EU-TMR Research Network ENDEAR (FMRX-CT98-0238).

+ Contact information: Jordi Brandts, Instituto de Análisis Económico (CSIC), Barcelona (brandts@cc.uab.es); Gary Charness, University of California, Santa Barbara (charness@econ.ucsb.edu).

1. INTRODUCTION

Notions of fair process and honorable behavior have potentially important implications for social and economic interactions, such as may be found in negotiations or organizations. Negotiators often have incentives to mislead others and private information may present the opportunity for doing so. Managers may be tempted to use deception to improve the chances of a desired response from a group of employees. However, there may be significant limitations to this kind of behavior: Where the character of interaction is highly interpersonal in nature, one must take into account the potential impact of social considerations on one's motivation, as this may impose bounds on selfish or dishonest behavior. We feel that it is a natural intuition that *deception* will be considered inappropriate behavior and may lead to substantial punishment behavior by the deceived. The experiment we present allows us to subject this intuition to a rigorous test.

The reaction to deception may also be generally relevant in the context of the current process of formulating more accurate theoretical models of human motivation. The most common assumption in economics is that people only care about maximizing their own income. But a large body of research has shown that many people choose to sacrifice money in laboratory experiments. A number of recent formal models presume that people are also motivated by considerations of altruism, inequity and aspects of the process by which an allocation is reached; while people may nevertheless maximize their utility, one's own money is not the sole determinant of utility.

This paper reports experimental results in a game with communication and the possibility of costly retribution.[1] Our design uses a game with three stages. Figure 1 shows the central

building block of our experimental design; the two numbers in each cell below refer to the respective material payoffs for the row and column players.

Figure 1

	B1	**B2**
A1	**2, 2**	**6, 9**
A2	**2, 2**	**12, 3**

In the first stage, player A sends a costless and non-binding message to player B stating that she intends to play A1 or intends to play A2. In the second stage, the players then choose actions simultaneously and one of the cells of the Figure 1 matrix is reached. Note that each cell can be reached via two different message-action paths - the sender may play in accordance with her message or she may not. In the third stage, player B may have additional options: If the outcome cell (A2,B2) is reached, player B can accept the (12,3) outcome or change it to (2,2); if the outcome cell (A1,B2) is reached, player B can accept the (6,9) outcome or change it to (8,7). Our experimental design includes possible misrepresentation, and provides a retribution mechanism with monetary consequences. This allows a clear expression of one's objection to deception.

Previous studies have demonstrated that people are willing to punish unfair actions, even at a personal cost. Fehr and Gächter (2000) find that free riders in a public goods experiment are heavily punished, and that this leads to substantially higher levels of cooperation. Sefton, Shupp, and Walker (2000) also note that sanctions in a public goods experiment are quite effective in achieving stable group allocations. Boles, Croson, and Murninghan (2000) study ultimatum game behavior when the proposer and the responder may have private information concerning

the pie size or the outside (rejection) payoff, respectively, and can exchange written messages. In their repeated-game fixed-matching design, they find that revelation of deceptive claims did not substantially increase deceptive proposers' subsequent offers, but responders who learn they have been deceived are more likely to reject these subsequent offers. However, in this design one's reaction to deception is mixed in with possible strategic considerations, since it is common information that one will be paired with the same person for the duration of the experiment.

With the study of the game presented above, we wish to add to the analysis of the effects of deception in both organizational settings and strategic situations. Previous work by Anton (1990), Lewicki (1993), and Shapiro and Bies (1994) finds that deception or lying can lead to moral outrage, and damage ongoing organizational relationships. Schweitzer and Croson (1999) suggest that "deception in organizations represents a significant managerial challenge across a broad range of functional areas." Our study provides clear and direct evidence that the willingness to punish an unfair action depends not only on the payoff outcomes in the relevant options, but also on the process that has led to the choice at hand. We use a design in which people interact with each other at most once, minimizing strategic reasons for punishment. The punishment cell (A2,B2) can be reached after either an accurate (A2) or deceptive (A1) message; regardless of the message, player B faces the same possible payoffs and the same choice between outcomes in the (A2,B2) cell. The only difference is the veracity of the message, and the substantial differences that we observe in punishment rates indicates that this element of the process is quite important to people.

Recent economic models of utility include some social values and preferences in the analysis. These models can be categorized by whether the process by which an allocation is reached and the perceived intentions of other players are relevant to an individual's preferences and choices. Purely distributional models assert that while people may sacrifice money to reduce disparities in material payoffs, they are unconcerned with the process leading to these

payoffs. Yet there is evidence that this is psychologically incorrect. In an example from the field, Kitzmann and Emery (1993) study parental satisfaction in child custody disputes, finding that differences in fathers' overall satisfaction can be attributed to procedural factors.[2]

Sen (1997) proposes that "a person's preferences over *comprehensive* outcomes (including the choice process) must be distinguished from the conditional preferences over *culmination* outcomes given the acts of choice," where the expression 'culmination outcomes' refers to material outcomes. He contends that choice functions and preferences may be affected by considerations such as "the *identity* of the chooser, the *menu* over which choice is being made, and the relation of the particular *act* to behavioral social norms that constrain particular social actions." Blount (1995), Charness (1996) and Offerman (2002) find evidence of Sen's *chooser dependence*: In sequential experimental games, second-mover responses differ according to whether the choice set is believed to be determined by a self-interested player or by a random mechanism. Brandts and Solà (2001), Charness and Rabin (2002) and Falk, Fehr and Fischbacher (forthcoming) have demonstrated the relevance of Sen's *menu dependence*: First-mover foregone alternatives significantly affect the choice made by the second player.

The models of pure distribution (e.g., Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) do not reflect menu dependence or chooser dependence, while the reciprocity models (e.g., Dufwenberg and Kirchsteiger, 1998; Falk and Fischbacher, 1999; Charness and Rabin, 2002) reflect both of these concepts. People's evaluation of deception would seem to fit well into at least the spirit of this general framework. However, the models that formalize the notion that preferences depend on comprehensive outcomes all focus on the actual choices made from the feasible choice set, and they do not explicitly consider the effect on utility from statements about

intended play.[3] To the extent that self-serving lies affect utility independently of their effects on material payoffs, a new model is needed.[4]

There is a considerable theoretical and experimental literature on cheap talk (see Crawford 1998 for a survey). Typically the issues considered are conditions under which cheap talk is informative and effective in achieving a desirable equilibrium or equilibria, under the assumption of standard (i.e., self-interested pecuniary) preferences. Essentially, deception (and cheap talk generally) is evaluated only with respect to how the payoff choice sets are affected, and not as an action *per se*.

We think that a retributive reaction to deception is potentially an important aspect of preferences over comprehensive outcomes and we are not aware of any studies that specifically examine (costly) reactions to misleading messages in one-shot environments. We study behavior for two different response elicitation methods: The *strategy method* (contingent responses made at every possible decision node) or the more standard *direct-response* method (responses made only to actual choices made and presented to the responder).

Using both elicitation methods, we find that many people send misleading messages and that people do sacrifice money to punish and reward. Significantly, we see a negative reaction to self-serving deception in both cases: Punishment rates are indeed higher when there has been a deceitful A1 message. Interestingly, punishment rate levels with direct responses are roughly double those with contingent responses. Reward rates are also somewhat higher, as 25% vs. 17% of our participants reward an A1 move by transferring two payoff units to the sender, thereby coming out behind instead of ahead.

Our results suggest that managers and negotiators should be aware of possible negative consequences from deliberate misrepresentation. In many situations (incentive pay, promotions,

etc.), the truth is eventually learned and misrepresentations are exposed. It appears that deception may be a breach of behavioral social norms.

2. FAIR PROCESS, COMMUNICATION, AND MODELS OF UTILITY

In this section, we first review some relevant literature on fair process, deception, and cheap talk. We then discuss the extent to which these issues are treated in recent utility models, and explore the predictions of these models in our game.

2.1 Fair process and deception. Material outcomes and payoffs are certainly a crucial factor in determining preferences and decisions; however, there may be other influences. There is a large body of work on procedural justice that supports the premise that process satisfaction is an important ingredient of human motivation. Thibaut and Walker (1975) and Tyler (1988, 1990) have argued that relational issues may dominate definitions of justice and that procedural satisfaction may be as important as outcome satisfaction. In the context of the study of organizations, Kim and Mauborgne (1996) find that "the exercise of procedural justice inspires managers to go beyond the call of duty and engage in innovative actions … on behalf of the organization." In this sense, procedural justice can inspire organizational citizenship behavior. There is also evidence that perceived procedural unfairness can lead to retaliatory behavior. Robinson and Bennett (1995) examine how employees respond to violations of the psychological contract and find behavior such as stealing from the company or co-workers, wasting company resources, lying about hours worked, and wrongfully blaming co-workers for mistakes.

Perceptions of the fairness of the process are also important for resource allocation in markets, negotiations, and labor relations. Kahnemann, Knetsch, and Thaler (1986), Barrett-

Howard and Tyler (1986), and Bies, Tripp, and Neale (1993) find that procedural information influences judgments of market exploitations. In their survey on fairness in negotiations, Tripp, Sondak, and Bies (1995) suggest that the allocation of resources may be of less concern to individual agents than procedural and interactional fairness. Charness and Levine (2000) find that the perceived fairness of a layoff is highly dependent on the manner in which the layoff is implemented.

Many studies in business ethics and negotiation address the specific issue of deception and its effects on behavior. While a satisfied party is more likely to maintain a positive and productive relationship with others in the environment, violations in relationships can lead to negative affect or even moral outrage. This is particularly true in the case of lying in negotiations (Anton, 1990; Lewicki, 1983). Although some feel that deception is just part of negotiation "dance," others (e.g., Shapiro and Bies, 1994) believe that such behavior can destroy trust and cooperation in ongoing organizational relationships. Bies and Tripp (1995) suggest that the harm done to the relationship by lying may be irreversible. Schweitzer, Brodt, and Croson (1999) find 36 of 66 "union negotiators" punish deceptive "city negotiators" when the true state of affairs is revealed, and that the distinction between lies of omission and lies of commission is important.

Lewicki and Stark (1996) analyze subjects' evaluations of ethically-questionable negotiation tactics. They suggest that players' perceptions of the "game" being played may be important. If people expect lies and deception, these may not produce much of a negative response. In Roth and Murnighan (1982), disbelief of messages was common. The stakes involved may also affect expectations: Tanbrunsel (1998) finds that increased incentives lead to

more misrepresentation and that the greater the incentive one has to engage in misrepresentation, the more that she expects that an opponent will do so.

Romer (1996) discusses the effects of deception in a political economy context. In his analysis, the U.S. social security system was created as an entitlement program: Payroll taxes were bundled with an explicit promise of certain future transfers. The reason why it is politically very difficult to cut back benefits is that "the act of making, then breaking, a promise induces a taste for punishing the offender" (p. 199). In the political sphere the punishment would be expressed through people voting against those who proposed a reduction in benefits.

2.2 Cheap talk. To the extent that promises or statements are unenforceable, they can be considered to be a form of *cheap talk*. While a message does not necessarily convey any information, subsequent behavior may be affected if a message is considered credible. Crawford (1998) points out that, even under the assumption of standard preferences, there is typically a problem of multiple equilibria.[5] Traditional equilibrium refinements are not generally helpful in this selection process. Farrell (1993) proposes restrictions based on the plausibility of out-of-equilibrium messages and Rabin (1990) presents a non-equilibrium concept that combines a credibility restriction with an assumption that players maximize their expected payoffs given their beliefs. These devices yield substantial and plausible restrictions on behavior, but maximizing is still defined purely by own monetary payoffs.

Experimentally, many studies find that cheap talk is quite effective in achieving Pareto-efficient outcomes. This is particularly true when players' interests are largely (or completely) in alignment. Cooper, DeJong, Forsythe, and Ross (1992) observe a high degree of success in attaining the payoff-dominant equilibrium outcome in a coordination game. Charness (2000)

finds that one-sided announcements of intended Pareto-optimal (but risky) play are extremely effective in a Stag Hunt, despite the fact that such messages may be self-serving and hence potentially less trustworthy.

Cheap talk is less effective when interests are in conflict, as false signals may not be credible. Sell and Wilson (1997) allow participants in a public-goods game to announce their intended contributions for the next period. Most announcements promised a higher level of contribution than was actually made, although the rate of "lying" decreased when people could check on each other's behavior; communication only enhanced cooperation with verification. In a Prisoners' Dilemma game in Charness (2000), 90% of all players ignored announcements of intended cooperation.[6]

In our setting, cheap talk is not really a coordinating device as much as it is an indicator of an A player's perceptions of the relevant social norms and of B's understanding of the game structure. What message do people think will be best, and are their beliefs correct? Cheap talk could potentially serve to achieve higher total payoffs by encouraging a B2 play. However, since the (2,2) that would result from a B1 play can also be achieved after (A2,B2), there is little reason to play B1 and it is in fact a rare choice.[1] Nevertheless, an A player may have so little trust in a B player's "rationality" that it seems worthwhile to advertise an A1 play, and we shall see that deceptive A1 messages are common.

2.3 Existing models of social utility and our hypotheses. The literature reviewed in 2.1 provides many insights into the behavioral effects induced by unfair processes and deception.

[1] In a sense, this serves as an embedded rationality test. If we observed many people choosing B1, we would be concerned that many people misunderstood the structure of the game. We see that B1 was chosen only 5% of the time, but many A's apparently did not trust that B's would choose B2 after an A2 signal and therefore signaled A1.

To what extent are these effects addressed in current utility models? Recent models of social preferences can explain punishment (and perhaps even reward). However, differential rates of punishment, the primary focus of our study, can only be the result of procedural distaste, as the set of payoffs available to B is independent of the signal sent by A.

Material reward is an important component of utility in all of the models mentioned in the previous section. Yet, many experimental participants show signs that nonpecuniary concerns are relevant to their decisions. For example, in the classic ultimatum game (Güth, Schmittberger, and Schwarze, 1982), many people reject unfair proposals, instead choosing zero payoffs for both themselves and the proposers. Difference- or inequality-aversion models such as Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) assert that people dislike unequal payoffs, but are indifferent to the process leading to their choice between outcomes. These models predict possible punishment in our game, as some B's may prefer (2,2) to (12,3). However, the rate of punishment should not be affected by whether or not a false announcement has led to this choice, so that the act of deception should not affect player B's action after the (A2,B2) cell has been reached.

On the other hand, models of reciprocity preferences include the issue of *intention*, the motivation or goal behind an action made by a (self-interested) party. In these models, a reciprocal action is the behavioral response to an action that is perceived as potentially "unkind" or "kind". It may make a big difference whether a particular action was intended or not. If an action was intentional, it also matters *why* it was chosen. If one perceives that another person is being kind or unkind by choosing a particular action or strategy, one may respond by sacrificing money to help or hurt this other person. Beliefs about the beliefs of other players are important.

One might suppose that the intention to deceive someone would naturally be seen as being unkind and unfriendly. However, the reciprocity models measure intention by comparing the outcomes available as a result of another player's choice(s) to the full range of potential outcomes, given the alternatives available to the other player. So an act of deception is only seen in a negative light if it unfavorably influences the victim's material outcome possibility set. Rabin (1993) explicitly defines kindness in terms of the best and the worst material outcomes that could result from another player's strategy. Dufwenberg and Kirchsteiger (1998) modify and extend this model so as to be more applicable to sequential games; again, the kindness of an action is defined in relation to its effects on the range of feasible material outcomes.[7]

Two recent models combine preferences over the distribution of material payoffs with reciprocity-based preferences. Falk and Fischbacher (1999) present a theory combining reciprocity with concerns about relative payoffs. In this model, the degree to which player B perceives that a particular outcome has been chosen intentionally depends on the options available to A.[8] B's perceptions of player A's kindness and intention influence B's choice.

Charness and Rabin (2002) motivate personal financial sacrifice by combining classical utilitarianism with Rawlsian preferences: People like to increase the social surplus (the total material payoff), but care more about helping low-payoff people than high-payoff people. However, they will withdraw their willingness to sacrifice for these principles according to their beliefs about how others are not following them, and may also have a taste for punishment (lowering their own payoffs to hurt bad actors). There is no positive reciprocity in this model: A decent person is *supposed* to do the right thing and receives no reward for actually doing it.

None of the reciprocity models mentioned predict that the punishment rate at (A2,B2) will depend on the message about intended play. Suppose that a false A1 signal is seen as an

attempt to induce a B2 play; if player B chooses B2, she foregoes the (2,2) payoff from B1. Yet in Rabin (1993) and Dufwenberg and Kirchsteiger (1998), no player will prefer (2,2) to either (12,3) or (6,9) absent negative reciprocity, so a pre-emptive play of B1 should not be made. While the action A2 may trigger B's annoyance that her final choice is between (12,3) and (2,2), instead of being between (6,9) and (8,7), the degree of annoyance is independent of the message.

In Charness and Rabin (2002), it is also true that everyone prefers (12,3) to (2,2) if the punishment parameter *f* has not been activated by player A's misbehavior. B1 play should not be observed, and the arguments above also apply here.[9] In Falk and Fischbacher (1999), it is possible that B would prefer (2,2) to (12,3) purely due to distributive concerns, so a false A1 signal could be seen as having some effect on B play. Nevertheless, even if B preferred (2,2) to (12,3), she can still impose the (2,2) outcome by choosing to punish A. These models do not predict that punishment rates will depend on the process leading to the (12,3) outcome.[10]

The closest analysis is the Dufwenberg (forthcoming) *psychological marital investment game*, in which the 2nd-mover spouse has pecuniary incentives to defect from his or her earlier promise to stay married, but may be constrained by feelings of guilt from promise-breaking. These feelings are linked to the promiser's 2nd-order beliefs about the trust level of the other spouse. While punishment is not considered in this game, incorporating social preferences in the utility function would allow punishment choices based on the perceived degree of violated trust.

We can now formulate our hypotheses. The null hypothesis is simply given by the standard model of individualistic preferences and predicts no punishment, no reward and, hence, excludes any differences in punishment or reward rates. Our first alternative hypothesis, H1, posits that overall punishment rate will be significantly different from zero; it is consistent with all the social utility models discussed above:

H1: A significant proportion of B players will choose to convert (12,3) to (2,2) after A2B2 play.

The two remaining hypotheses go beyond the above models and capture process satisfaction aspects of social interaction. H2 formulates the differential punishment conjecture:

H2: The proportion of B players choosing to convert (12,3) to (2,2) will be greater after an A1 signal than after an A2 signal.

The predictions of the social utility models is somewhat mixed with respect to whether B may choose to reward an A1 play. The distributional models cannot easily accommodate such a choice.[11] The Charness and Rabin (2002) model does little better, as only a very high weight on the minimum payoff can explain reward. Since Rabin (1993), Dufwenberg and Kirchsteiger (1998), and Falk and Fischbacher (1999) allow positive reciprocity, the decision to reward can be explained for B's, but only with high reciprocity parameters. Nevertheless, we expected to see a sizable proportion of positive reciprocal responses after an A1 play. Accordingly, our third alternative hypothesis is:

H3: There will be a significant reward rate for A1B2 play.

We have no hypothesis concerning differences in reward rates across messages.

3. EXPERIMENTAL DESIGN AND PROCEDURES

We conducted our experimental sessions at UC-Berkeley. A total of 212 people participated in exactly one of the eight sessions. Each unit of experimental payoffs was set equal to $1.50; average earnings were around $16, including a $5 show-up fee, for the one-hour session. Recruiting was conducted primarily through the use of campus e-mail lists. An e-mail message that was sent to randomly-selected people through the Colleges of Letters, Arts, and Sciences provided the bulk of the participants, so our sessions typically included individuals

from a broader range of academic disciplines than is common in economics experiments. Since the vast majority of Berkeley students use e-mail, selection bias from this recruiting method should be minimal, at least with respect to other laboratory experiments. Instructions are provided in the Appendix.

People met in a large classroom that was divided into two sides. Individuals sat at non-adjacent desks where instruction packages had been placed. The subjects on opposite sides of the room had different roles (A or B), with each person on one side of the room paired with one person on the other side of the room

As mentioned in Section 1, there are 3 stages in our game: The first is the announcement stage, the second is the simultaneous choice stage and the third is the retribution stage. In the first stage, each person in the sender role sends an announcement about his intended play to an anonymous receiver: the announcement is a non-binding statement about which choice, A1 or A2, the sender will make in the second stage. In the second stage, after the message has been transmitted, both players simultaneously choose actions. The sender chooses between A1 and A2 and the receiver chooses between B1 and B2. In the third stage, the receiver has an option to change the payoffs if she has played B2. If (A1, B2) has been chosen, she can give the sender 2 units and so change the payoffs to (8,7). If (A2, B2) is the outcome, she can change the payoffs from (12,3) to (2,2). Matched players can reach the (A2, B2) cell by two message-action paths - one where the message has been A1 and the other where the message has been A2.[12]

If money is the only element in one's utility function, the only subgame-perfect Nash equilibrium involves actions (A2, B2) without punishment or reward in the third stage; the message is irrelevant. Two other Nash-equilibrium-strategy combinations, both of which

presume there is nonpecuniary utility, consist of the sender playing A1 and B punishing A2, and 1) not rewarding A1 or 2) rewarding A2.

The experimental design and the payoff calibration are motivated by several considerations. First, we wanted a simple environment, where the (binary) choices and associated payoffs were transparent to the players. Another issue relates to our main objective of having an environment in which it is plausible to expect self-serving lies. In our game we expected many senders' preferred outcome to be (12,3) and, therefore, their preferred action in the simultaneous choice stage to be A2. However, a sender may be concerned that a receiver will choose B1 after an A2 message, and so may send an A1 message to encourage a B2 choice, but then actually play A2.[13]

Given that the punishment payoffs in the retribution stage of the game are (2,2), there is really no obvious reason for the receiver to choose B1 in the action stage. However, there may exist plausible explanations for receivers choosing B1; the fact that it is not completely transparent why the sender should expect a B1 choice after an A2 message does not interfere with the analysis we wish to perform. As long as we obtain sufficient observations of the (A2, B2) cell being reached after A1 and A2 announcements, we can compare receivers' behavior in the two cases.

Having the punishment payoffs be the same as the B1 payoffs is particularly useful with respect to the utility models discussed earlier, since allowing B to return to the B1 payoffs means that B's payoff range is unaffected by whether an A1 message is false. In a sense, this permits us to isolate the effect of the deception *per se*, without respect to its influence on the payoffs available. If we had allowed the B1 payoffs to be different than the punishment payoffs, the pure effect from deception would have been confounded by possible effects from a change in the

payoff range induced by a false signal. Nevertheless, other B1 payoffs could have been chosen. We wanted to have (A1,B1) and (A2,B1) payoffs identical so that an unhappy player B could unilaterally determine the result in the game. The payoff combination we chose was calibrated to yield a sizable proportion of both false statements by A and B2 choices.[14]

Following Bolton, Brandts, and Katok (2000) and Brandts and Charness (2000), each game was played twice, so that each person was a sender once and a receiver once. Participants were assured that no two people were ever paired twice, and were not informed about the final outcome of the first play of the game when they made their decisions for the second play. This feature allows us to obtain data from one-shot interactions and also permits us to examine whether subjects play in a consistent manner across roles.[15] If the decision to punish deception is not influenced by whether one has sent (or intends to send) a false message, punishment could be seen as not being based on a consistent behavioral norm. Following the two periods, a coin was tossed to determine the period used for actual payment.

The choice of response-elicitation method is an important issue in experimental economics, as there is a tension between efficiency in data–gathering and the quality of the data. We collected data for two different response-elicitation methods. In our first five sessions, receivers were not told the decisions actually made by the senders before they were asked for their choices of whether to punish and reward. Instead each receiver (who knew the message she had received) was asked to designate (after his B2 play) a contingent choice if the sender actually played A1 and a contingent choice if the sender actually played A2.[16] One obvious advantage of this approach is that we can obtain a full set of two responses regardless of the sender's play.

This *strategy method* (Selten, 1967) plausibly induces different behavior than does the standard "direct-response" method. Roth (1995) mentions on pg. 323 that "having to submit entire strategies forces participants to think about each information set in a different way than if they could primarily concentrate on those information sets that may arise in the course of the game." This method may, hence, capture more reflective behavior. In contrast, there is clearly a certain element of immediacy to receiving information about an actual choice. One might expect that some actions would trigger stronger emotional responses in such an environment. Many social psychologists feel that visceral elements have a strong effect on behavior. For example, Loewenstein (1996) observes that such factors may cause people to be "out of control" and act against their own self-interest.

Since is quite plausible that punishment behavior may be different when responders are presented with actual A1 or A2 play, we tested whether a behavioral effect from deception is also observed under more visceral conditions.[17] We conducted three additional sessions, in which each responder was told the choice actually made by the first mover, and made a direct response to this choice and only to this choice. The design in these sessions was otherwise identical to that in the initial five sessions.

4. RESULTS

4.1 Strategy-method sessions. Tables 1 and 2 present a summary of the data from the strategy-method sessions involving 118 subjects, where the notation s(A1) and s(A2) refers to A1 and A2 messages, respectively. By punishment we refer to the receiver choosing payoffs (2,2) after an (A2,B2) realization and by reward to the choice of (8,7) after (A1,B2). We remind the reader that

each receiver was asked to designate separate choices for the case where the sender actually played A1 and the case where the sender actually played A2.[18]

Table 1 – Strategy-method Play in Stages 1 and 2

	A		**B**	
	Play A1	**Play A2**	**Play B1**	**Play B2**
s(A1)	51/76 (67%)	25/76 (33%)	4/76 (5%)	72/76 (95%)
s(A2)	11/42 (26%)	31/42 (74%)	2/42 (5%)	40/42 (95%)

Most A players (76/118, or 64.4%) choose an A1 message. We find that about 1/3 (25/76) of all A1 messages are false, as they are followed by A2 play. Overall, nearly 70% of all participants play in accordance with the message they sent. Perhaps reassuringly, 95% of the subjects played B2. If one interprets the choice of B1 as a pure mistake, note that a 5% "error rate" is not at all unusual in laboratory experiments. Note that the same proportion of B players played B2 whether the message was A1 or A2.

Table 2 – B's Punishment and Reward Rates (Strategy Method)

	s(A1)	**s(A2)**	***Z*-stat**		**s(A1)**	**s(A2)**	***Z*-stat**
Punish A2	19/69 (28%)	5/40 (12%)	1.83	**Reward A1**	10/72 (14%)	9/39 (23%)	-1.23

We use the test of the equality of proportions (Glasnapp and Poggio, 1985) to test for significance. The *Z*-statistic reported is the normal approximation to the binary distribution, defined by the difference between the proportions divided by the standard error of the difference.

The overall punishment rate, 24/109, is significantly different from zero ($Z = 5.19$, $p < .0001$), rejecting the predictions of the standard model in favor of H1.[19]

The rate of punishment of action A2 is substantially and significantly higher given a false signal. As we have a directional hypothesis regarding the punishment rates, we use a one-tailed test and find the difference significant at $p = .034$.[20] This is support for the general notion that the process by which an outcome is reached affects its ultimate attractiveness; specifically, we reject the predictions of the null in favor of H2.[21] A deceptive message is not seen to be appropriate by many participants and a substantial number of these are even willing to sacrifice money to express their displeasure.

We also find that, overall, 19 out of 111 subjects, over 17%, choose to reward a play of A1. This evidence is favorable to our third alternative hypothesis, since this is quite significantly different from 0% ($Z = 4.58$, $p < .0001$). We also feel that the 17% reward rate is much higher than what would have been obtained if we simply gave B players unilateral power to choose (6,9) or (8,7) and had no messages, as it is rare for an experimental participant to elect to receive less than another person when a choice to receive more is available. As we do not run this treatment, we cannot make strong claims, but this is nevertheless suggestive of some kind of positive reciprocity.[22]

4.2 Direct-response sessions. Tables 3 and 4 present a summary of the data from the direct-response sessions involving 94 subjects. The results in stages 1 and 2 are similar to those in the strategy-method sessions. A slightly higher proportion (69 of 94, or 73.4%) of A's choose an A1 message and a slightly lower proportion (16/69) of all A1 messages are false. The *Z*-statistics for these comparisons are 1.40 and 1.30, respectively, so that the differences are not statistically

significant. Once again, nearly all (98%, $Z = 1.12$, n.s., for the strategy-method comparison) of the B subjects played B2, with a slightly higher rate of B1 play after an A2 message.

Table 3 – Direct-response Play in Stages 1 and 2

	A		**B**	
	Play A1	**Play A2**	**Play B1**	**Play B2**
s(A1)	53/69 (77%)	16/69 (23%)	0/69 (0%)	69/69 (100%)
s(A2)	3/25 (12%)	22/25 (88%)	2/25 (8%)	23/25 (92%)

Table 4 – B's Punishment and Reward Rates (Direct Response)

	s(A1)	**s(A2)**	***Z*-stat**		**s(A1)**	**s(A2)**	***Z*-stat**
Punish A2	9/16 (56%)	6/22 (27%)	1.80	**Reward A1**	13/53 (25%)	0/3 (0%)	0.98

The overall punishment rate, 15/38, is again significant ($Z = 4.32$, $p < .0001$). Behavior in these sessions also shows a strong effect from deception, with punishment rates still about twice as high after an A1 signal and A2 play. Using the one-tailed test, the difference is significant at $p = .036$, a very similar value to the one for the strategy-method sessions.[23] However, these punishment rates were approximately double those found using the strategy method. Although we conclude that the effect of deception is robust to elicitation method, there does indeed appear to be an effect on the level of the punishment rate (and perhaps the reward rate) from a more visceral environment.

There is an overall reward rate of 23% (13 of 56), which is significantly different from 0% ($Z = 3.83$, $p < .0001$). This rate is slightly higher than that in the strategy-method sessions, but the difference is not significant ($Z = 0.95$). There is no significant difference in reward rates across messages, although the test is weak since there are very few subjects who choose A1 after an A2 message. In summary, our data are consistent with our three alternative hypotheses.

4.3 Role-rotation and consistency. Since we use role rotation in our experimental design, we can examine whether people exhibit a "consistent attitude" towards punishment and reward behavior. By this we mean that people who punish a false message or reward an A1 play are less likely to be deceptive and are more likely to play A1. Table 5 illustrates this internal consistency:

Table 5 – Consistency across Roles (Pooled data)

	Punish Lie As B	**Don't Punish Lie as B**	**Reward A1 as B**	**Don't Reward A1 as B**
Lie as A	3/28 (11%)	25/28 (89%)	4/34 (12%)	30/34 (88%)
Don't Lie as A	32/79 (41%)	47/79 (59%)	28/133 (21%)	105/133 (79%)

There is a strong relationship between whether one lies as A and one's willingness to punish a lie. Those people who do not lie in the A role are nearly four times as likely to punish in the B role. The *Z*-statistic for the difference of proportions is 2.89 ($p = .002$, one-tailed test) for punishment, showing a behavioral consistency with respect to punishment. This "internal consistency" is an indication that subjects understand the situation, and parallels results such as

Kahneman, Knetsch, and Thaler (1986), where 88% chose to punish a selfish chooser (sacrificing $1) if they had not themselves been selfish choosers, but only 31% punished behavior they would engage in themselves.

On the other hand, the correlation between lies and rewards is not as strong. Since these are very different acts, perhaps this should not be surprising. People who lie as A are less likely to reward than those who don't lie, but the difference is only slight, and at best marginally significant ($Z = 1.23$, $p = .109$, one-tailed test).

4.4 *Ex post* profitability. How do the effective levels of retribution affect players' material payoffs? Table 6 shows expected material payoffs for all combinations of message and action.[24] If a sender only cares about her own material payoff she should signal A2 and then play A2; it does not pay to send an A1 message prior to an A2 play.[25]

Table 6 – Ex post expected material payoffs (Strategy Method)

Message, action	Sender expected payoff	Receiver expected payoff	Combined expected payoff
s(A2), A2	10.36	2.84	13.20
s(A1), A2	8.86	2.69	11.55
s(A2), A1	6.23	8.25	14.48
s(A1), A1	6.05	8.38	14.43

On the other hand, a sender with a sufficient positive weight on the receiver's material payoff should play A1, as this leads to higher social benefits than A2, regardless of the signal. We see (Table 1) that 53% of senders play A1, giving them lower expected material payoffs *ex post* than from an A2 play. While we cannot know precisely why this choice was made (perhaps an exaggerated fear of punishment), the results suggest this may reflect pro-social behavior.[26]

Table 7 – Ex post expected material payoffs (Direct Response)

Message, action	Sender expected payoff	Receiver expected payoff	Combined expected payoff
s(A2), A2	8.69	2.67	11.36
s(A1), A2	6.38	2.44	8.82
s(A2), A1	5.52	8.44	13.96
s(A1), A1	6.49	8.51	15.00

Ex post, A could choose (A,B) payoffs of (8.69, 2.67) or (6.49, 8.51). The latter is more efficient, but A may not wish to sacrifice 2.10 to increase B's payoff by 5.84. Note that the response elicitation method has very little effect on the orderings of the expected payoffs.

5. DISCUSSION

Previous studies indicate that process satisfaction is a highly salient consideration in many environments. The negotiation and business ethics literature suggests that deception, while a fairly common practice, often induces negative responses. Generally, people react in an adverse manner when they feel their right to fair treatment has been violated. In accordance with this view, we find that false messages lead to punishment much more frequently than do accurate messages, even though the choices between culmination outcomes is the same in both cases; this holds for both the strategy and the direct-response elicitation methods. People appear to have a *per se* objection to deception, even when it does not affect their actions, material payoffs, or future relationships.

In addition, while responses to favorable play are not our primary focus, we do find that many people donate money in response, choosing to come out behind; this suggests the presence of some kind of positive reciprocity. We also find that a message receiver's decision whether to

punish deception or reward a favorable play is significantly correlated with her behavior as a message sender. This is evidence that the decision to punish is part of a consistent value orientation rather than being simply arbitrary or random.

While our results provide a clean illustration (with monetary incentives) of the consequences of deception, the effect may be a bit surprising, as false messages are explicitly permitted in the instructions and there is no obvious deterrent value for punishment. Since a deceived person who plays B2 can always obtain the material payoff available from playing B1, deception does not reduce that person's choice set. Therefore, it seems that people simply do not like being misled as such and that this triggers a taste for punishment, which may be related to self-respect issues.

Punishment has been defined (Kadzin, 1975) as the presentation of an adverse event or the removal of a positive event following a response. Romer (1996) suggests two reasons that people may choose to threaten (and impose) punishment. Threats of punishment may have strategic value (even in the static case) or actual punishment may have future deterrent value (dynamic case). A person may also have a taste for punishing others that is triggered by a sequence of events. In this case, a person punishes because it is satisfying to do so. Given the non-repeated nature of our design, there should be no deterrence motivation or future financial considerations.[27]

Throughout the paper, we have implicitly assumed that lies will invariably be detected. While we feel that the true state of affairs is often revealed eventually in negotiations and organizations, there may be situations where the probability of detection is low. In such cases, the probability of punishment would be correspondingly lower, so that deception might well be strategically effective. However, the credibility of a deceptive act might also be reduced if the

would-be deceived party knows that there is only a small probability of detecting deception. In addition, the use of punishment itself may depend on material payoffs. Zwick and Chen (1998) find that the higher the "price" of fairness, the lower the demand for it. We might also expect less punishment where it is not as cost-effective.

Our experiment is an instance of the influence of intention – promises and deception. This is an important issue in potentially cooperative negotiations and business environments, where trust can lead to mutually beneficial outcomes. Are breaches of trust perceived by some to be *per se* violations of the Sen (1997) behavioral social norms?[28] In this case, one's distaste for being played the fool can be enough to generate the urge to punish the transgressor, even if no money is lost. While this seems difficult to model, perhaps one could tie the degree of disappointment from a breach to the change in expectations (or the reference level) induced by a false promise. This could relate to the *let-down aversion* postulated by Dufwenberg and Gneezy (2000).

Our results have implications for theories of human motivation in economic contexts. All current models of social utility predict some punishment for A2 play. However, we have seen that these models do not predict the differences in punishment rates observed in our experimental game. While this is not surprising for models that do not reflect the process by which an outcome has been reached, it points out a blind spot in models that try to address process satisfaction by only considering the intentions of other entities.

Our work has value for organizational decision-makers as it combines insights from psychology and economics and suggests that process satisfaction is important in economic environments. Managers should be cautious about the use of deception, as the people they supervise may react adversely upon discovery of the deceit. Violations of implicit codes of

conduct cause displeasure, even if there is no clear financial harm. In the field, there may well be ongoing relationships among people who interact, so that an immediate negative reaction may have far-reaching adverse effects.

REFERENCES

Anton, R., "Drawing the Line: An Exploratory Test of Ethical Behavior in Negotiations," 1990, The International Journal of Conflict Management, **1**, 265-280.

Barrett-Howard, E. and T. Tyler, "Procedural Justice as a Criterion in Allocation Decisions," 1986, Journal of Personality and Social Psychology, **50**, 296-304.

Bies, R. and T. Tripp, "Beyond Distrust: 'Getting Even' and the Need for Revenge," 1995, in Trust and Organizations, R. Kramer and T. Tyler, eds., Sage: Newbury Park, CA.

Bies, R., T. Tripp, and M. Neale, "Procedural Fairness and Profit Seeking: The Perceived Legitimacy of Market Exploitation," 1993, Journal of Behavioral Decision Making, **6**, 243-256.

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

Boles, T., R. Croson, and K. Murninghan, "Deception and Retribution in Repeated Ultimatum Bargaining," 2000, Organizational Behavior and Human Decision Processes, **83**, 235-259.

Bolton, G. and A. Ockenfels, "ERC: A Theory of Equity, Reciprocity, and Competition," 2000, American Economic Review, **90**, 166-193.

Bolton, G., J. Brandts, and E. Katok, "How Strategy Sensitive are Contributions? A Test of Six Hypotheses in a Two-Person Dilemma Game," 2000, Economic Theory, **15**, 367-387.

Brandts, J. and G. Charness, "Hot vs. Cold: Sequential Responses in Simple Experimental Games," 2000, Experimental Economics, **2**, 227-238.

Brandts, J. and C. Solà, "Reference Points and Negative Reciprocity in Simple Sequential Games," 2001, Games and Economic Behavior, **36**, 138-157.

Charness, G., "Attribution and Reciprocity in an Experimental Labor Market," 1996, mimeo.

Charness, G., "Self-serving Cheap Talk and Credibility: A Test of Aumann's Conjecture," 2000, Games and Economic Behavior, 33, 176-94.

Charness, G. and D. Levine, "When are Layoffs Acceptable? Evidence from a Quasi-Experiment," 2000, Industrial and Labor Relations Review, **53**, 381-400.

Charness, G. and M. Rabin, "Understanding Social Preferences with Simple Tests," 2002, Quarterly Journal of Economics, **117**, 817-869.

Charness, G. and M. Rabin, "Expressed Preferences and Play in Simple Games," 2001, mimeo.

Cooper, R., D. DeJong, R. Forsythe and T. Ross, "Communication in Coordination Games," 1992, Quarterly Journal of Economics, **107**, 739-771.

Crawford, V., "A Survey of Experiments on Communication via Cheap Talk," 1998, Journal of Economic Theory, **78**, 286-298.

Dawes, R., J. McTavish and H. Shaklee, "Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation," Journal of Personality and Social Psychology, **35**, 1-11.

Dufwenberg, M., "Marital Investments, Time Consistency, and Emotions," forthcoming in Journal of Economic Behavior and Organization.

Dufwenberg, M. and U. Gneezy, "Measuring Beliefs in an Experimental Lost wallet Game," 2000, Games and Economic Behavior, **30**, 163-82.

Dufwenberg, M. and G. Kirchsteiger, "A Theory of Sequential Reciprocity," 1998, mimeo.

Falk, A. and U. Fischbacher, "A Theory of Reciprocity," 1999, mimeo.

Falk, A., E. Fehr, and U. Fischbacher, "On the Nature of Fair Behavior," forthcoming in Economic Inquiry.

Farrell, J., "Meaning and credibility in cheap-talk games," 1993, Games and Economic Behavior, **5**, 514-531.

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

Fehr, E. and K. Schmidt, "A Theory of Fairness, Competition, and Cooperation," 1999, Quarterly Journal of Economics, **114**, 817-868.

Glasnapp, D. and J. Poggio, Essentials of Statistical Analysis for the Behavioral Sciences, 1985, Merrill: Columbus.

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

Kadzin, A., The Token Economy, 1977, Plenum Press: New York.

Kahnemann, D., J. Knetsch, and R. Thaler, "Fairness as a Constraint on Profit-Seeking: Entitlements in the Market," 1986, American Economic Review, **76**, 728-741.

Kitzmann, K. and R. Emery, "Procedural Justice and Parents' Satisfaction in a Field Study of Child Custody Dispute Resolution," 1993, Law and Human Behavior, **17**, 553-567.

Kim, W. and R. Mauborgne, "Procedural Justice and Managers' In-role and Extra-role Behavior: The Case of the Multinationals," 1996, Management Science, **42**, 499-515.

Lewicki, R., "Lying and Deception: A Behavioral Model," 1983, in Negotiating in Organizations, M. Bazerman and R. Lewicki, eds., Sage: Beverly Hills.

Lewicki, R. and N. Stark, "What is Ethically Appropriate in Negotiations: An Empirical Examination of Bargaining," 1996, Social Justice Research, **9**, 69-95.

Loewenstein, G., "Out of Control: Visceral Influences on Behavior," 1996, Organizational Behavior and Human Decision Processes, **65**, 272-92.

Offerman, Theo, "Hurting Hurts More than Helping Helps," 20002, European Economic Review, **46**, 1423-1437

Rabin, M., "Communication between Rational Agents," 1990, Journal of Economic Theory, **51**, 144-170.

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

Rabin, M., "Psychology and Economics," 1998, Journal of Economic Literature, **36**, 11-46.

Robinson, S., and R. Bennett, "A Typology of Deviant Workplace Behaviors: A Multidimensional Scaling Study, 1995, Academy of Management Journal, **38**, 555-572.

Romer, P., "Preferences, Promises, and the Politics of Entitlement," 1996, in Individual and Social Responsibility: Child Care, Education, Medical Care, and Long-Term Care in America, V. Fuchs, ed., University of Chicago Press: Chicago.

Roth, A., "Bargaining Experiments," in Handbook of Experimental Economics, J. Kagel and A. Roth, eds., 1995, 253-348.

Roth, A. and J. Murnighan, "The Role of Information in Bargaining: An Experimental Study," 1982, Econometrica, **50**, 1123-1142.

Schotter, A., K. Weigelt, and C. Wilson, "A Laboratory Investigation of Multiperson Rationality and Presentation Effects," 1994, Games and Economic Behavior, **6**, 445-468.

Schweitzer, M. and R. Croson, "Curtailing Deception: The Impact of Direct Questions on Lies and Deceit," 1999, mimeo.

Schweitzer, M., S. Brodt, and R. Croson, "Deception, Distance, and Technology: A Comparison of Videoconference and Telephone Negotiations," 1999, mimeo.

Sefton, M., R. Shupp, and J. Walker, "The Effect of Rewards and Sanctions in Provision of Public Goods," 2000, mimeo.

Sell, J. and R. Wilson, " 'Liar, Liar …,': Cheap Talk and Reputation in Repeated Public Goods Settings," 1997, Journal of Conflict Resolution, **41**, 695-717.

Selten, R., "Die Strategiemethode zur Erforschung des Eingeschränkt Rationalen Verhaltens im Rahmen eines Oligopolexperiments," 1967, in Beiträge zur Experimentellen Wirtschaftsforschung, H. Sauermann, ed., 136-168.

Sen. A., "Maximization and the Act of Choice," 1997, Econometrica, **65**, 745-779.

Shapiro, D. and R. Bies, "Threats, Bluffs, and Disclaimers in Negotiations," 1994, Organizational Behavior and Human Decision Processes, **60**, 14-35.

Tanbrunsel, A., "Misrepresentation and Expectations of Misrepresentation in an Ethical Dilemma: The Role of Incentives and Temptation,"1998, Academy of Management Journal, **41**, 330-339.

Thibaut, J., and L. Walker, Procedural Justice: A Psychological Analysis,1975, Erlbaum: Hillsdale, NY.

Tripp, T., H. Sondak, and R. Bies, "Justice as Rationality: A Relational Perspective on Fairness in Negotiations," 1995, in Research on Negotiation in Organizations, Volume 5, R. Bies, R. Lewicki, and B. Sheppard, eds., 45-64, JAI Press: Greenwich, CT.

Tyler, T., "What is Procedural Justice?: Criteria Used by Citizens to Assess the Fairness of Legal Procedures," 1988, Law Soc. Review, **22**, 301-355.

Tyler, T., Why People Obey the Law: Procedural Justice, Legitimacy, and Compliance, 1990, Yale University: New Haven.

Zwick, R. and X. Chen, "What Price Fairness? A Bargaining Study," 1998, Management Science, **44**, 119-141.

NOTES

[1] Webster's Third New International Dictionary defines retribution as "the dispensing or awarding of punishment or reward according to the deserts of the individual."

[2] Couples were randomly assigned to mediate or litigate their child custody dispute. In nearly every case, physical custody was awarded to the mother. Yet fathers were far more satisfied with the mediation process.

[3] Rabin (1998) states that "people determine their dispositions towards others according to motives attributed to these others, not solely according to actions taken." Models such as Rabin (1993), Dufwenberg and Kirchsteiger (1998), Falk and Fischbacher (1999), and Charness and Rabin (2002), discussed further in section 2.3, consider the decisions and motivations of other agents to be important determinants of behavior. A key common element in these models is that perceived intentions are important: *Why* did another player make a particular choice? What was he trying to achieve and how appropriate is this goal?

[4] Dufwenberg (forthcoming), discussed in section 2.3, does study some aspects of the breaking of promises.

[5] For example, in any cheap-talk environment there is always a "babbling equilibrium" in which messages are ignored.

[6] On the other hand, Dawes, McTavish, and Shaklee (1977) find that "relevant" (discussion about the situation at hand) face-to-face communication leads to a substantially higher rate of cooperation in a commons dilemma situation.

[7] We thank Matthew Rabin and Martin Dufwenberg for helpful comments in this regard.

[8] For example, if player A has only two options (say 1 and 2), and each of these gives player B a choice between (8,2) or (0,0), A's choice of option 1 would not be considered intentional or unkind.

[9] There is a caveat, however: *f* is a free parameter, so that one could (in principle) consider the taste for punishment to be sensitive to the simple act of deception. But this is not explicitly modeled.

[10] It is possible that if B strongly expects an A2 play and has a self-control problem, she might wish to tie her hands by playing B1. However, the issue of self-control is far beyond the scope of these models.

[11] It is easy to prove that changing the payoffs from (6,9) to (8,7) is incompatible with the parameter constraints in Fehr and Schmidt (1999). Bolton and Ockenfels (2000) do not provide a functional form, so no firm conclusions can be reached. However, the loss of 2 units of pay would be compared with a very mild improvement in relative payoffs, so that this choice could only occur with a rather extreme degree of difference aversion.

[12] There is a second possibility for false messages in our design. The sender may announce A2 and then choose A1, behaving more favorably to the receiver than announced. A possible rationale for this behavior is to surprise the receiver in order to elicit the reward. As will be shown below, we observed numerous instances of this behavior.

[13] It is true that a B player who has chosen B2 because of an A1 signal may wish to punish an A2 play and, hence, make A's lie useless. However, some A's may believe that B's will not always choose to punish. In our one-shot environment, we do not expect A's beliefs and B's beliefs to be correct.

[14] We did not make reward symmetric with punishment; suppose B could reward A by changing the (6,9) payoffs to (16,8). This social payoff could be so attractive that we might not observe much deception. In any event, our aim here was not to determine the respective strengths of the impulses to reward or punish, but rather to see if some B's would choose to reward an A1 play even when it was relatively unattractive to do so. Thus, we made reward a 1-1 transfer from player B to player A, such that the direction of their relative payoffs is reversed. No additional social surplus is generated by a reward, and voluntarily relegating one's self to the lower material payoff seems psychologically unappealing.

[15] Charness and Rabin (2002) provide evidence that this role-reversal approach does not appear to affect behavior in simple games, in 19 games tested using both with and without role reversal.

[16] We chose not to elicit contingent responses to messages, as we were concerned that asking people to first envision a hypothetical message and then consider their hypothetical responses to this hypothetical message might be pushing the strategy method too far. Double hypothetical responses are considerably more abstract and complex, and the risk of confusing the experimental participants seemed unacceptably high. In addition, handing a person a slip of paper marked in pen may be helpful in establishing experimental credibility.

[17] However, evidence from simple experimental games has suggested that the elicitation method does not significantly affect responses. Brandts and Charness (2000) specifically study this question in two 2-person binary-choice games (Prisoners' Dilemma and Chicken) and find no significant difference in behavior due to elicitation method. Schotter, Weigelt, and Wilson (1994) examine a simple game involving two choices per player. Holding the representation of the game (as a tree) constant, they find no difference between the strategy and direct-response elicitation methods. Nevertheless, these studies do not involve reactions to dishonest behavior.

[18] While each B player thus made two choices, each statistical test used contains only one of these choices (B's action if A1 was chosen or if A2 was chosen). Since these tests are run separately, there is no problem of multiple observations per participant.

[19] While there should be 112 total responses to A1 messages and to A2 messages (118 subjects less those 6 who played B1), 3 subjects failed to indicate a response to a hypothetical A2 play and one subject failed to indicate a response to a hypothetical A1 play. Thus, there are only 109 total responses to A2 play and 111 total responses to A1 play.

[20] It is standard practice in hypothesis testing to use a one-tailed test when there is an ex ante directional hypothesis. See Siegel and Castellan (1988, p. 8), for example.

[21] An argument could be made that the true difference in rates is higher, as the punishments after an A2 play and an A2 message may include some B's who would have played B1, except for the small chance of an A1 play.

[22] Interestingly, the reward rate is more than 50% higher when A1 is played after an A2 message, although the difference is not statistically significant with the small *N*. Perhaps people respond more favorably when the sender shows a favorable change of heart.

[23] If we pool the rejection data across elicitation methods, the difference in punishment rates after A1 and A2 signals is significant at p = .020 (one-tailed test) or .040 (two-tailed test).

[24] Given receivers' behavior, it appears that material payoffs account only for part of their motivation. The other part stems from the satisfaction of punishing a self-serving lie and of rewarding a pleasant surprise. It is conceivable that senders' payoffs are also affected by a non-material component, although this force is more difficult to gauge; senders' may derive some satisfaction from sending a truthful message.

[25] This result suggests that in a dynamic relationship, false messages would be driven out, since they ultimately are not in the sender's interest. However, our focus here is on the study of preferences. For this purpose what matters are the reactions to false statements, when they are made.

[26] We can examine the economic effects of the punishment of false messages by comparing the expected payoffs shown in the first and second rows of Table 3 reveal. If one relates the reductions of sender and receiver payoffs to the largest possible reduction from the payoffs shown in the first row of table 3 down to a payoff of 2, one obtains a fraction of 1.5/8.36 for the sender and of .15/.84 for the receiver. Both fractions are close to 18%.

[27] It is also possible that people are willing to provide a socially-beneficial "object lesson" to deter further anti-social behavior outside the laboratory.

[28] While deception could be construed to be an act that violates a behavioral social norm, Sen (1997) does not include this factor in the model presented. He writes (p. 751) that "Some types of influences of choice acts are more easy to formalize than others, and these include: (i) chooser dependence, and (ii) menu dependence."