Can Observers Predict Trustworthiness?*

Michèle Belot†, V. Bhaskar‡, and Jeroen van de Ven§

August 5, 2008

Abstract

We analyze experimental evidence on whether untrained subjects can predict how trustworthy an individual is. Two players on a TV show play a high stakes prisoner's dilemma with pre-play communication. Our subjects report probabilistic beliefs that each player cooperates, before and after communication. Subjects correctly predict that women, and players who voluntarily promise that they will cooperate, are more likely to cooperate. They are also able to distinguish truth from lies when a player is asked about his or her intentions by the host. In consequence, and in contrast with the psychology literature, our naive subjects are able to distinguish defectors from cooperators, with the latter inducing beliefs that are 7 percentage points higher. We also study Bayesian updating in the natural and complex context, and find mean reversion in beliefs, and reject the martingale property.

JEL Classification Numbers: C72, C93, D64, D83.

Keywords: trust, promises, Bayesian updating, detecting deception, martingale property of beliefs.

*Thanks to Shane Frederick for providing us with the cognitive ability test, to the University of Essex for financial support and CREED for use of their subject database. We are grateful for comments from Joao Santos Silva and various seminar audiences.

†Department of Economics, University of Essex, Wivenhoe Park, CO4 3SQ Colchester, United Kingdom. E-mail: mbelot@essex.ac.uk

‡Department of Economics, University College London, Gower St. WC1E 6BT London, United Kingdom. Email: v.bhaskar@ucl.ac.uk

§Department of Economics, University of Amsterdam, ACLE, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands, E-mail: j.vandeven@uva.nl

He that has eyes to see and ears to hear may convince himself that no mortal can keep a secret. If his lips are silent, he chatters with his fingertips; betrayal oozes out of him at every pore.

Sigmund Freud (1905).

1 Introduction

Economic and social relationships cannot be governed entirely by formal contracts, leaving ample scope for opportunistic behavior. This highlights the importance of trust and trustworthiness in sustaining efficient economic transactions. Knack and Keefer (1997) use cross-country data and find that an increase in trust, as measured by attitudinal surveys, is related to higher economic growth. La Porta et al. (1997) find that higher levels of trust are associated with less corruption and higher judicial efficiency.

Just as important as the absolute level of trustworthiness or trust, is the extent to which an individual can predict how trustworthy his or her partner in a transaction is likely to be. Numerous experiments show that individuals are very heterogeneous in terms of trustworthiness (e.g. Fehr and Schmidt, 1999; Blanco et al., 2006). If "betrayal oozes out of every pore", then many profitable transactions can be undertaken, even if the overall level of trustworthiness is not very high. Trustworthy agents will also have an advantage in economic and social transactions. Conversely, if individuals find it hard to distinguish agents with different propensities, then one-shot transactions will necessarily be inefficient and one must rely either on repeated interaction or contractual mechanisms to deter opportunism. This suggests another reason why trust and trustworthiness is higher between pairs of individuals who are socially closer (see Glaeser et al., 2000), since they may be more able to "read" one another. Indeed, there is a strand of evolutionary game theory which argues that if propensities are even partially observable, then non-maximizing behavior, such as being trustworthy,

may be evolutionarily stable (Frank, 1988; Guth and Yaari, 1992; Dekel et al., 2007). The *involuntary truth-telling hypothesis* (Frank, 1988; Ockenfels and Selten, 2000) asserts that types are indeed partially observable – opportunists inadvertently look and behave differently from trustworthy people, despite their attempts at deception.

The practical importance of detecting opportunism is quantifiable in sectors of the economy that are particularly susceptible to fraud, such as insurance, the tax and benefit system and the criminal justice system. These sectors spend large sums of money in training individuals so that they can distinguish fraudulent claims from genuine ones. Anderson (1999) estimates that fraudulent transfers amount to $550 billion annually in the US. Laband and Sophocleus (1992) report that investments in white-collar crime prevention cost $216 million in the US in 1985. However, we should stress that the scope of our enquiry is wider than that of fraud. In many business relationships, opportunism is not illegal, and the consequent moral or social sanctions are weaker.

Despite the economic importance of the question, is trustworthiness distinguishable from opportunism, there is very little work on the subject. The closest seems to be work by psychologists on whether experimental subjects can distinguish true statements from false ones – we discuss this work in the following section. One practical problem is that opportunism and deception are by their intrinsic nature hard to observe: successful deception often goes unnoticed and honesty is rarely completely verifiable in the field. This paper attempts to overcome this problem. We use DVDs of a game show where two players play a prisoners' dilemma game and must choose between cooperating and defecting. Our experimental subjects watch the show up to the point that these decisions are made, and are asked to predict behavior. This has two advantages. First, the players on the game show freely choose their decisions, so that they incur any psychological or moral costs associated with opportunism. Second, we observe perfectly whether these players are opportunistic or trustworthy. Compared to

a fully experimental design, using the game show has the further advantage that the stakes for the players are high, giving real incentives to appear trustworthy.

More concretely, our experimental subjects are shown episodes of a television game show,[1] at the end of which two players play a prisoner's dilemma game. Each player must choose whether to share (S) or to grab (G) a sum of money X. The monetary payoffs to the row player as a function of his own action and that of his opponent, the column player, are depicted in Fig. 1.

	S	G
S	$\frac{1}{2}X$	0
G	X	0

Figure 1: Monetary Payoffs

That is, if both players share, they each get $\frac{X}{2}$; if only one player shares, his opponent gets the entire amount; and if both players choose to grab, they both get zero. The median stake value of X is €1,683, so that the stakes are substantial. Prior to playing the prisoner's dilemma, the two players communicate, allowing them to convey the honesty and sincerity of their intentions. Our experimental subjects are shown the behavior of the game show players until the crucial decision, and asked to report their beliefs, i.e. the probability they assign to a player deciding to share. Subjects make a prediction regarding any player twice. A subject's *interim belief* is her prediction made before she observes the communication stage. Her *final belief* is her prediction after communication. This allows us to see how subjects update their beliefs. That is, we are able to study (Bayesian) updating in the context of a natural and complex problem, trustworthiness. In addition, we also elicit each subject's *prior belief*

[1] The show was broadcast in the Netherlands under the name 'Deelt ie 't of deelt ie 't niet', which translates as 'Does s(he) share or not?'.

or her *base rate*, i.e. her prediction of the average rate of cooperation across all episodes of the game show.

Our empirical methodology is based on the random assignment of episodes to groups of subjects. Each subject makes predictions for several players, with differing characteristics and behavior. We use the within-subject variation in predictions to identify how subjects update their beliefs in response to perceived signals. More specifically, we regress subject beliefs on player characteristics and behavior, and compare the estimated coefficients with the analysis of the game show data, which shows how these same variables affect sharing behavior. Behavior on the game show is analyzed in a companion paper (Belot et al., 2006).

The main findings are as follows. Subjects appear to use the right cues forming interim beliefs – for instance, they rightly believe that a woman is more likely to share than a man, although they underestimate the magnitude of the difference. Subjects also ignore characteristics such as age or attractiveness that our analysis shows to be irrelevant.

Our most interesting findings relate to how subjects update after observing communication by the players. The game show data shows that a player's explicit promise to share is associated with a significantly higher probability of sharing by the player, when this promise is made voluntarily, i.e. at the player's own initiative. The experimental subjects correctly revise their final beliefs upwards on observing a *voluntary promise*, although they underestimate the size of this effect. Thus talk is not cheap when voluntarily undertaken, and is correctly perceived as such by our observers. Some of the players who make a voluntary promise are in fact lying, and we find that our subjects are unable to distinguish truth from lies. Thus the involuntary truth-telling hypothesis is not valid for voluntary promises.

Our second set of findings relate to the response made by a player when he (or she) is asked explicitly and unexpectedly about his intentions by the

presenter of the show. Such a player invariably replies that he will share. These *elicited promises* are not associated with any greater propensity to share in our data. However, we find evidence of the "Columbo effect"[2] – our subjects are able to read these responses and distinguish truth from lies. They revise their final beliefs upwards in response to elicited promises that in fact turn out to be true, but not in the case of elicited promises that turn out to be false. Thus the involuntary truth-telling hypothesis seems to be valid when the question is asked unexpectedly.

By using these signals of cooperation, our subjects can distinguish cooperators from defectors – the average player who shares induces subject beliefs that are 7 percentage points higher than the average player who grabs.[3] This estimate is significant and also large in relation to much of the existing literature, which suggests that predictions by non-professional subjects do not perform significantly better than "chance".

Our final set of findings relate to Bayesian updating, since we are able to study how subjects update in a complex and natural setting, where they are exposed to a range of different signals. We present and test a simple Bayesian model, and find that our basic results reported above are robust to allowing for non-linear effects that arise in a Bayesian setting. However, we also find significant violations of Bayesian rationality. Most significantly, we reject the martingale property of beliefs, that the prior should equal the average posterior. Instead, we have mean reversion in beliefs – subjects with a low prior tend to have a higher posterior, while those with a high prior have a lower posterior.

The rest of the paper is organized as follows. Section 2 discusses the related literature. Section 3 presents the data and experimental set-up. In section 4, we present results based on a linear specification showing subjects perceive player

[2] Named after the TV detective Columbo who finishes interviewing suspects, and then invariably surprises them with "one last question". We are grateful to Steve Nickell for this analogy.

[3] This estimate is based on a regression of subject predictions upon an indicator for the sharing decision, with subject fixed effects and random effects.

characteristics and communication. Section 5 presents the theory and evidence regarding Bayesian updating. Section 6 analyzes how subject characteristics affect their beliefs and discusses the determinants of how accurate subjects are, and the final section concludes.

2 Related Literature

Economists have recently become interested in the act of deception – Gneezy (2005) investigates the considerations that affect deception, while Wang et al. (2006) find that pupils of subjects dilate when they send deceptive messages. The flip side of this question, whether deception is detectable, has been the preserve mainly of psychologists. In the typical psychology experiment, "actors" are instructed to lie or to tell the truth, and an observer assigns a truth value to the actor's statement.[4] Ekman and Friesen (1974) showed nurses a movie that was either pleasant or nasty; in both cases the nurse was instructed to tell an interviewer that the movie was pleasant. These interviews were videotaped, and showed to observers. Observers are given little or no information on how the statements have been selected, so that they may have little basis for forming a "prior", regarding the underlying probability of lying. Observers are generally not paid for accurate predictions, and the subjects telling lies are usually not paid for successful deception, although the nurses were told that this was important for their career.[5] The general finding is that observers are not very accurate in detecting lies. Their success rate in classifying statements is usually not significantly higher than 50%. DePaulo et al. (1985, p. 327), conclude that "deception accuracy usually exceeds chance, although rarely by an impressive margin." However, some groups of people with training, such as federal officers and secret service agents, manage to do better (Ekman et al. (1999)).

It may be argued that the participants in these experiments have been in-

[4] Ekman (1985) and DePaulo and Friedman (1998) provide good surveys.

[5] Exceptions are Frank and Ekman (1997) and Kraut and Poe (1980), who gave a bonus to participants who were judged to be honest.

structed to lie, reducing their feelings of guilt and possibly making detection harder. In an interesting study, Mann et al. (2002) show police officers videotapes of statements by real criminal suspects. Here again, the officers were provided no information on how these statements were selected, giving little basis for forming a prior regarding the underlying rate of lying. It is therefore hard to tell whether optimistic predictions are due to an optimistic prior or to the interpretation of signals.

On predicting trustworthiness, previous work includes Frank et al. (1993) and Brosig (2002), who let subjects communicate before playing a prisoner's dilemma game, and ask them to predict the decision of their opponents. Their focus is on the overall accuracy of predictions rather than the cues used and perceived by subjects. They argue that subjects are able to predict their opponent's play with an accuracy rate above chance; however, their attribution of what constitutes "chance" when the true cooperation rate differs from 50% is somewhat questionable.[6] Ockenfels and Selten (2000) conduct a bargaining experiment where subjects are randomly assigned high or low bargaining costs, and this is private information. They find that observers are able to guess the true costs of the bargainers 55% of the time; however, this success in guessing above chance is explained by objective features, such as the length of the bargaining negotiations, rather than involuntary signals.

The question addressed by the present paper is broader than lie detection, since players do not have to lie on our show. However, some players do declare their intention to share, and this allows us to see whether our subjects can detect lies specifically. The main methodological difference in comparison with the previous literature is three-fold. First, we show a random sample of TV episodes to our subjects, so that we do not select for a certain proportion of

[6] If the subjective cooperation rate perceived by an observer exceeds 0.5 and the true cooperation rate is $p > 0.5$, then the observer would score p by simply predicting cooperation every time. On this definition of "chance", the observers in these studies do not do better than chance.

cooperators (or truth-tellers). Having done so, we explicitly elicit prior beliefs, interim beliefs and final beliefs from the subjects, so that we can study how subjects update their beliefs in response to the signals they receive. For example, we can study how the subjects revise their beliefs according to the gender or age of the players or the statements they make. Secondly, the decisions of the players on the TV show and their statements are entirely voluntary. By contrast, in many of the psychology experiments, individuals are instructed to lie or tell the truth. A final difference is that we elicit probabilistic beliefs rather than a 0-1 prediction. This allows us to avoid problems with predictions when the underlying cooperation rate is biased away from 0.5, for in this case, it is often optimal to make extreme predictions such as always predicting share for all episodes they observe.

It may be useful to contrast our methodology with standard experiments, e.g. those by Ekman and Friesen (1974) or Mann, Vrij and Bull (2004), where an observer watches an individual, and has to make a binary report on whether this individual is telling the truth or lying. In a Bayesian framework, the observer's belief depends not only on the signals he perceives (i.e. what he sees in the videotape or the interview) but also upon his prior belief. One difficulty is that the observer may not have any basis for forming a prior belief. In the nurses experiment (Ekman and Friesen), the sample was chosen so that nurses were telling the truth half the time. In the police interview experiment, the tapes had a mixture of true and false statements, but in both studies the subjects were unaware of the basis of selection.

Making a binary report rather than a belief also has the disadvantage that the subject does not know how to choose a threshold value of his belief for this decision. In real life, even if the decision is binary (e.g. whether to investigate further a police suspect or not), the threshold value depends upon the costs and benefits associated with each alternative and will not in general be 0.5.

A further problem arises in determining what a good prediction rate is (i.e. one that is better than chance) when the underlying cooperation rate differs from 0.5. Suppose that the true cooperation rate is $p > 0.5$ and the prior μ_i also exceeds 0.5. In this case, an observer whose signals are uninformative can achieve a success rate of p by simply predicting cooperation for every player.[7] Finally, the observer only reports a different belief if the signal is such that it makes the posterior cross the threshold. Binary predictions are therefore likely not to pick up signals very well. These problems are avoided by asking observers to report beliefs, rather than coarser information. To summarize, the advantages of our approach are as follows:

1. We ask subjects to report probabilistic beliefs rather than a binary variable, and provide incentives such that reporting true beliefs is optimal.

2. Each subject sees a random sample of episodes. Subjects could therefore have a prior based on how likely it is that players share in such a large stakes environment on TV.

3. Each subject makes predictions for a sample of players, either eight or twelve in number. This allows us to identify how subjects vary their beliefs depending upon player characteristics, by using the within-subject variation in predictions. For the communication stage, we can go further, by only using the *change* in predictions made by the subject for the same player, before and after communication.

4. We have a large number of predictions, over 3000.

[7] The literature (e.g. Frank et al. (1993) or Brosig (2002)) defines the "chance prediction rate" as $pq_i + (1-p)(1-q_i)$ where q_i is the proportion of cooperative predictions by observer i. When $p > 0.5$, this is always less than p, the success rate that is achievable by the strategy of always predicting cooperation. Subjects do not usually follow such strategies, suggesting in turn that they may not be maximizing their probability of being correct.

3 Background and description of the data

3.1 The game show

We showed our subjects episodes of the game show 'Does s(he) share or not?', broadcast in the Netherlands in 2002. The game starts with five players, who accumulate earnings by answering quiz questions. At the end of each of three rounds, one player is eliminated by the player with the highest score of that round. At the end of the third round, the player with the highest earnings chooses one of the two other remaining players. These are the two finalists, and the eliminated players play no further role in the game. The total prize money X is the combined earnings of the two finalists, who then play the prisoner's dilemma game by simultaneously deciding to share or grab. The two finalists first make speeches to each other. This communication is free format.

3.2 Experimental set-up

We ran a total of seventeen sessions, 5 of them in Utrecht in May 2006 and 12 in Amsterdam in May 2007, with a total of 169 subjects (89 in Utrecht and 80 in Amsterdam). The subjects were mainly social science students, with one session using support and support staff from the university of Utrecht. The 69 episodes were randomly assigned to sessions, the typical subject seeing either four episodes (Utrecht) or six episodes (Amsterdam).

The game show lasts for 25 minutes, consisting of several rounds of quiz questions, and the sequential elimination of players until only two of the original five remain. Since most of the show may not be relevant to predicting the decisions made by the final two players, and in order to economize on time, we showed our subjects shortened episodes. In the 2006 sessions, subjects saw two entire episodes (lasting around 25 minutes each) and two shortened episodes (6 minutes each). The shortened versions did not include the rounds with the quiz questions, but include the stage where the players introduce themselves,

and the stage where one of the three remaining players is to be eliminated. Since an analysis of the data confirmed that watching the entire game show did not make a material difference to the predictions made by subjects, in the 2007 sessions, we did not show the entire show, but only showed the subjects four shortened episodes, and two slightly longer (medium) episodes, which also included the last round of quiz questions. This variation in the length of episodes also provides different "treatments", allowing us to see how length and the quantity of information affects predictions.[8]

The show was paused several times, at which points we asked subjects to make predictions (see Fig. 2). First right before the selection decision is made, where we asked subjects to predict which player would be chosen for the final.[9] The other two occasions the show was paused, subjects were asked to indicate the probability of sharing for each of the two players in an episode — before the players made speeches to each other, and after – which allows us to infer the effects of the content of communication upon subject beliefs.[10] Subject choices were restricted to a discrete grid, with step size 0.1, i.e. a subject was asked to choose $p \in \{0, 0.1, ..., 0.9, 1\}$. Subjects were paid according to a quadratic scoring rule, and were told that they should report their beliefs truthfully in order to maximize expected earnings. The experiment lasted about one hour and a half, and subjects earned approximately €18, including a €4 participation fee.[11] Subjects were not given any feedback on the actual decisions made by players, so as to prevent learning.

We also ascertained various personal details of the subjects and also asked

[8] A full description of the nature of these different treatments is provided in the appendix.

[9] These selection predictions are used for another project (see Belot et al. (2006)) and we do not discuss them further here.

[10] In the Utrecht sessions, the show was only paused twice: right before the selection decision, and after the communication. Because the opponent of the lead player was still unknown to the subjects at the selection decision, we asked them to predict the sharing probability of the lead player against each of the two possible opponents. See the Appendix for details.

[11] The non-student session was run mainly with support staff. To provide real incentives to them, we offered them a lunch for participation, and out of the 14 we randomly chose two subjects who earned around €40, depending on their choices. We found no differences in their predictions as compared to the student groups.

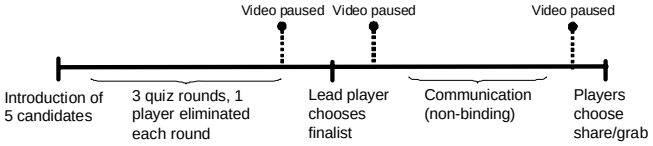

Figure 2: Timeline game show

them to estimate the average cooperation rate over all the 69 episodes of the show, with a payoff of €1 if their prediction lay within 5 percentage points of the true value. Anonymity of answers and earnings was guaranteed. We also asked the subjects to play the prisoner's dilemma game of the game show. In Utrecht, this play was hypothetical – we asked the subject what he or she would do as a player on the game show. In Amsterdam, each subject was asked to make a choice (share or grab). Two subjects were selected randomly afterwards, and their choices were implemented, with a total stake of €200. We asked the Amsterdam subjects if they would like to donate a part of their earnings to a charity, Warchild.[12] The Amsterdam subjects also did a short test of cognitive ability, taken from Frederick (2005). We also conducted a small experiment to infer their risk preferences, by asking them to make several decisions, where in each case a subject had to choose between a fixed amount (€2) and a lottery with varying stakes. The instructions given to subjects can be found in the Appendix.

4 Perceived cooperative traits

The players on the game show share 43 percent of the time (see Table 1). This is exactly what the subjects expect on average when asked to report the average

[12] Warchild raises funds to help the children victims of war across the world and is among the best known charities in the Netherlands. The subjects could donate 0, 5, 10 or 20% of their earnings.

Table 1 - Sharing probabilities - Summary statistics

	Mean	sd
Actual game show data ($N = 138$)	.43	.50
Prior ($N = 169$)	.43	.21
All predictions before communication ($N = 1,672$)	.52	.29
All predictions after communication ($N = 1,672$)	.52	.28

sharing probability across all 69 episodes. However, their average prediction is 9% higher, at 52%. This suggests that a typical subject judges the median player that she sees to be above the median in terms of trustworthiness. That is, on seeing any specific individual, a subject is more likely to trust this individual than when asked the question in abstract, perhaps because of the positive image the players give of themselves.

4.1 Cooperative cues

We now investigate how subjects form their beliefs and, in particular, how they update their prior based on the public information they receive about the players. Our empirical strategy is based on the random assignment of episodes to groups of subjects, and on the panel aspect of the data. Indeed, subjects have been asked to report predictions for multiple players with different observable characteristics. We can therefore identify precisely how subjects update their beliefs based on the available information, using the within-subject variation only. This may be compared with analysis of actual behavior on the game show (see Belot et al. (2006)).

We consider the predictions reported after the communication stage,[13] as a function of observable characteristics, excluding communication – section 4.2 investigates the effects of communication on the updating of beliefs. We model the prediction of subject i regarding player j, p_{ij} , as follows:

[13] Very similar results are obtained if we use predictions made before communication.

$$p_{ij} = \alpha_i + \beta' X_j + \delta_j + \varepsilon_{ij}, \tag{1}$$

where α_i is a subject fixed effect, X_j is a vector of observable characteristics of player j, and the error term includes a player-specific component (δ_j) and an idiosyncratic i.i.d. component (ε_{ij}). β can be interpreted as the average update in response to a signal X.

Column 1 in Table 2 reports the results from the analysis based on the data from the game show (see Belot et al. (2006)). Women are almost 20 percent more likely to share than men[14], and those who contribute relatively little to the final prize money are more likely to share. A larger stake slightly increases the sharing probability. Age and attractiveness, on the other hand, have statistically insignificant effects. Column 2 shows how these characteristics determine the predictions reported by subjects. We find that subjects pick up some of the relevant cooperative signals, but tend to underestimate the magnitude of the effects. They believe that women are more cooperative and they also expect a positive relationship between the size of the stakes and cooperativeness. On the other hand, they do not perceive a correlation with the relative contribution to the prize, although there is a strong relationship in the actual data. Finally, they correctly do not associate age or attractiveness with cooperativeness.

4.2 Communication

We now turn to the perception of promises and lies. Before the players make their final decision in the last round, they get the opportunity to make a brief speech. This speech is "cheap talk" in the sense that any statements made are not binding and do not affect monetary payoffs. In contrast with other

[14] Such gender effects are not new. See Belot et al. (2006) for a discussion how they relate to the existing literature.

Table 2 - Actual cues and perceived cues

Dependent variable	Actual realization		Beliefs	
	(1)		(2)	
Female	.19	(.09)**	.08	(.03)***
Age	.00	(.01)	.00	(.001)
Contribution	-.72	(.31)**	-.04	(.05)
Prize (x €1,000)	.03	(.01)**	.01	(.003)***
Attractiveness	-.03	(.07)	-.02	(.02)
N observations	138		1672	
R-squared			.23	

Standard errors are reported in parentheses, *, ** and *** correspond to 10%, 5% and 1% significance levels respectively

experiments, the statements made by the players are entirely voluntary, and no one has been instructed to lie.

46% of the players make an explicit promise to share, i.e. they specifically state "I will share" or "I promise to share". Those who do not make an explicit promise usually talk about what they intend to do with the money; try to convince the other player to share, or say in general terms that "sharing is good". We also make a distinction between a voluntary promise and one that is made in response to an explicit question by the compere of the show, who sometimes explicitly asks a player whether he or she will share. We call such a promise an "elicited promise" – indeed, with only one exception all contestants replied "yes" to the compere's question. 13% of our contestants make an elicited promise. Those who make a voluntary promise are almost 50 percent more likely to share than those who don't. Elicited promises, on the other hand, are uncorrelated with actual behavior.

Since we ask subjects to report predictions for the same players before and after the communication stage (denoted p_{ij}^{before} and p_{ij}^{after} respectively), we can identify precisely the effect of communication on predictions. In particular, we can see how subjects revise their beliefs upwards for players who make an explicit promise. We estimate the following equation:

$$p_{ij}^{after} - p_{ij}^{before} = \beta_0 + \beta_1 promise_j + \varepsilon_{ij}, \quad (2)$$

where β_0 and β_1 are constants, $promise_j$ is a dummy indicating whether the player made an explicit promise or not and ε_{ij} is an i.i.d. random disturbance term. We consider different types of promises; and distinguish between truthful promises and lies.

First, we find that subjects do see that promises are correlated with cooperative behavior, and are to some extent capable of identifying liars. Table 3 reports the results of a regression, where the dependent variable is the *change in beliefs* of the subjects, as a function of the content of communication by the player. The first column shows that subjects increase their beliefs (about the probability of sharing) when a player make a promise, although the effect is only significant for voluntary promises. The effect is relatively small though - predictions increase by about 5% for voluntary promises in comparison to the average increase of 50% in the game show data. So, overall, subjects fail to capture the magnitude of the effect of this strong predictor of cooperativeness.

Since some players who make a promise choose to grab, column 2 investigates whether subjects are able to identify liars. The results are quite striking. While subjects do not see any significant difference in cooperativeness across players who make a voluntary promise, irrespective of whether they are liars or not, they identify liars very clearly when the promise has been elicited. There could be several reasons for this. *Those who choose to initiate the promise* may be better liars or lies prompted by surprise may be harder to disguise. In any case, subjects become substantially more optimistic regarding players from whom a promise has been elicited and who will indeed cooperate.[15]

Because communication is sequential (the lead player speaks first), there

[15] A related result by Charness and Dufwenberg (2007) is that the form and content of a promise is important in a form of a trust game: personalized promises are effective, while impersonal bare promise messages are not. Interestingly, they find impersonal promises have no effect on beliefs, while personalized promised do change beliefs of players.

Table 3 - Communication and beliefs

Dependent variable	Difference in predictions before and after communication			
	(1)		(2)	
Voluntary promise	.05	(.01)***	.04	(.02)**
Elicited promise	.02	(.02)	.13	(.05)***
Voluntary promise & lying			.03	(.02)
Elicited promise & lying			-.15	(.05)**
Constant	-.03	(.01)***	.11	(.10)
R-squared	.008		.16	
N. obs	1672		1672	

Standard errors are reported in parentheses, *, ** and *** correspond to 10%, 5% and 1% significance levels respectively

might be differences in how subjects update their beliefs depending on whether the player is first or second to speak. In the actual data, we found no correlation between the player's speech and the behavior of the opponent. For example, we found no evidence that the second player's speech or behavior depends on whether the first player makes an explicit promise or not. We investigated whether there were any systematic differences in the predictions made for players 1 and 2. The results remain essentially the same and are not reported for the sake of brevity. However, interestingly, we do find a significant correlation (.25) in predictions made for both players. That is, subjects believe that the decisions to cooperate will be correlated across players.

4.3 Overall quality of predictions

We now consider how well our subjects are able to predict the sharing decision overall. To this end, we regress final beliefs upon the sharing decision, while including subject fixed effects and random effects. We find that if a player shares, this is associated with a 7 percentage points increase in subject beliefs. Thus our subjects are able to use cooperative cues reasonably well in making predictions.

5 Bayesian updating

We now consider how updating by our subjects corresponds to a standard Bayesian model. First, we investigate the possible non-linear effects that arise in Bayesian updating. Second, we test the martingale property of beliefs.

5.1 Non-linear updating

Let us consider a player who will make a decision $\omega \in \{G, S\}$. G corresponds to grab whereas S corresponds to share. The observer, or subject i, has to report the probability that $\omega = S$. This is based on observing signals emanating from the player. We model the subject as a Bayesian, who makes his prediction based upon his prior and upon the signals that he observes. Let us suppose that the decision maker starts with a subjective prior belief μ_i that a player will share. For the purposes of this section, we will use the term prior to include either a) the subject's interim belief, or b) her estimate of the average sharing probability across all episodes (i.e. what is termed the prior in the rest of the paper). The posterior would then correspond to a) the final belief or, b) the interim belief or final belief, respectively.

Suppose now that the subject observes a signal, σ, which takes values in a finite set. Let $q_i(s|\omega)$ denote the probability assigned by the subject to the signal taking value s given that the state equals ω. The decision maker's posterior belief that $\omega = S$ is given by $\pi_i(s)$:

$$\begin{aligned} \pi_i(s) &= \frac{\mu_i q_i(s|S)}{\mu_i q_i(s|S) + (1-\mu_i) q_i(s|G)} \\ &= \frac{\mu_i \ell_i(s)}{\mu_i \ell_i(s) + (1-\mu_i)}, \end{aligned} \tag{3}$$

where $\ell_i(s) = q_i(s|S)/q_i(s|G)$ is the subjective likelihood ratio for signal s. This is a non-linear function of the prior and the likelihood ratio. However, algebraic manipulation of this expression yields the following linear specification:

$$\ln \frac{\pi_i(s)}{1-\pi_i(s)} = \ln \frac{\mu_i}{1-\mu_i} + \ln l_i(s). \tag{4}$$

In other words, if we transform variables so that the dependent variable is constructed from posterior and prior beliefs using the above formula, this should be a linear function of indicator variables corresponding to the various signal realizations.

Consider first the case where the prior is the subject's reported average sharing rate across all episodes. In this case, the dependent variable can be constructed from the posterior (i.e. the reported interim or final belief), and we can control for the prior by including subject fixed effects in the regression. Of course, this assumes that different individuals have the same subjective likelihood ratio when evaluating signals. Second, one may wish to study how subjects update from interim to final beliefs, for any given player. In this case, our dependent variable is constructed as $\ln \frac{\pi_i(s)}{1-\pi_i(s)} - \ln \frac{\mu_i}{1-\mu_i}$. In either case, the specification is now linear as a function of signal realizations.

We therefore estimate the following model:

$$\ln \frac{p_{ij}}{1-p_{ij}} = \alpha_i + \beta' X_j + \delta_j + \varepsilon_{ij}. \tag{5}$$

For the characteristics (k) that are binary dummy variables (such as gender, promises), the estimated coefficient $\widehat{\beta}_k$ equals $\ln \frac{(p_{ij}|X_{kj}=1)}{1-(p_{ij}|X_{kj}=1)} - \ln \frac{(p_{ij}|X_{kj}=0)}{1-(p_{ij}|X_{kj}=0)}$, i.e. to the log likelihood ratio. We can calculate the corresponding value in the actual data:

$$\ln \frac{(\widehat{p}_{ij}|X_{kj}=1)}{1-(\widehat{p}_{ij}|X_{kj}=1)} - \ln \frac{(\widehat{p}_{ij}|X_{kj}=0)}{1-(\widehat{p}_{ij}|X_{kj}=0)}, \tag{6}$$

where $\widehat{p}_{ij}$ is the predicted value of $\widehat{p}_{ij}$ conditional on X_k and on the average values of all other characteristics included in the vector X. To ease the exposition, we report results using dummy variables for all player characteristics. Table 4 shows the results. The results confirm what we have found with

Table 4 - Actual and perceived cues (log likelihood ratios)

	Actual data		Final beliefs	
Female	.78	(.25)***	.62	(.17)***
Age > 32	.14	(.12)	.14	(.17)
30%≤Contribution ≤ 70%	-1.17	(.66)*	-.02	(.22)
Contribution>70%	-.92	(3.05)	-.08	(.27)
Score > average score	-2.09	(3.61)	-.35	(.17)*
Attractiveness > 4	.18	(.26)	-.13	(.19)
Voluntary promise based on cross-section	2.12	(.33)***	.21	(.17)
Voluntary promise based on difference before/after communic.			.36	(.10)***

Standard errors are reported in parentheses, *, ** and *** correspond to 10%, 5% and 1% significance levels respectively

the linear specification. Overall, subjects do perceive the correct signals, but underestimate their magnitude. However, for gender, we cannot reject that the update is fully consistent with Bayesian updating.

A second implication of Bayesian updating from equation (3) is that for given subjective likelihood ratios, the extent of updating is a non-linear function of the prior, and the subject updates more for non-extremal priors, and less for extreme priors. However, we might also expect that this relationship is overturned if more extreme priors are also correlated with more extreme likelihood ratios. That is, a subject who is cautious may be simultaneously be more likely to have prior beliefs that are closer to 0.5 and also have likelihood ratios that are close to one.

To investigate more closely this non-linearity, we differentiate between three ranges of priors ([0,0.2], [0.3,0.7] and [0.8,1]), where the prior is the subject's estimate of the probability of sharing across all 138 players on the game show. We now allow estimate a linear specification, while allowing the coefficient on characteristics to vary depending upon the prior. These results are reported in Table 5. We find little support for non-linear updating. For most variables, we cannot reject that the coefficients are identical across priors, and for gender we find that those with a low prior update their beliefs significantly more than those with priors in the medium range. This suggests that those with more extreme

Table 5 - Actual cues and perceived cues depending on the prior

Dependent variable	Actual realization		Beliefs	
	(1)		(2)	
Female	.19	(.09)**	.066	(.027)**
Female & low prior			.062	(.034)*
Female & high prior			.023	(.063)
Age	.00	(.01)	.001	(.002)
Age & low prior			-.001	(.002)
Age & high prior			-.001	(.004)
Contribution	-.72	(.31)**	-.045	(.055)
Contribution & low prior			-.017	(.072)
Contribution & high prior			.097	(.129)
Prize (x e1,000)	.03	(.01)**	.009	(.004)
Prize & low prior			-.005	(.007)
Prize & high prior			.008	(.008)
Attractiveness	-.03	(.07)	-.012	(.020)
Attractiveness & low prior			-.027	(.028)
Attractiveness & high prior			.007	(.053)
N observations	138		1664	
R-squared			.24	

Standard errors are reported in parentheses, *, ** and *** correspond to 10%, 5% and 1% significance levels respectively

priors also have more extreme likelihood ratios. Similar results are obtained when we consider the difference between final and interim beliefs. There is no evidence that the magnitude of updating is greater for interim beliefs that are intermediate rather than extreme.[16]

5.2 Testing the martingale property of beliefs

We now test another implication of Bayesian updating, that of the martingale property of beliefs: the prior must equal the weighted average of posteriors, where the weights depend upon the probabilities of the signal realizations. The prior can be written as:

$$\mu_i = \sum_s \Pr(s)\pi_i(s), \tag{7}$$

[16] For reasons of space we do not report econometric results on this, but this may be verified by inspecting figure 4 which follows later in this section.

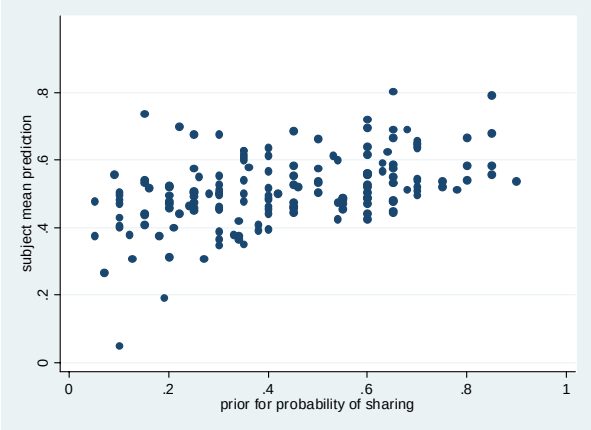

Figure 3: Prior and subject mean prediction

where $\Pr(s)$ is the subjective probability that signal s will be observed:

$$\Pr(s) = \mu_i q_i(s|S) + (1 - \mu_i) q_i(s|G). \tag{8}$$

That is, for any prior of any subject, the prior equals the expected value of the posteriors. The empirical implication is that the average of realized posteriors should equal the prior.

Figure 3 plots the actual subject's mean of predicted final beliefs against the subject's estimate of the average probability of sharing across all 138 players on the game show, i.e. her prior. At low values of the prior, the mean posterior is higher than the prior, but at high values, the mean posterior is lower than the prior.[17] In other words, we have mean reversion in beliefs, and a violation of the martingale property.

The hypothesis of equality of priors and average posteriors can also be tested by comparing interim beliefs (i.e. predictions before communication) and final

[17] A t-test shows that the posterior is systematically higher than the prior for values of the prior values below .6 and is systematically lower than the prior for values of the prior above .6. Averaged across all priors, the mean final belief is greater than the mean prior.

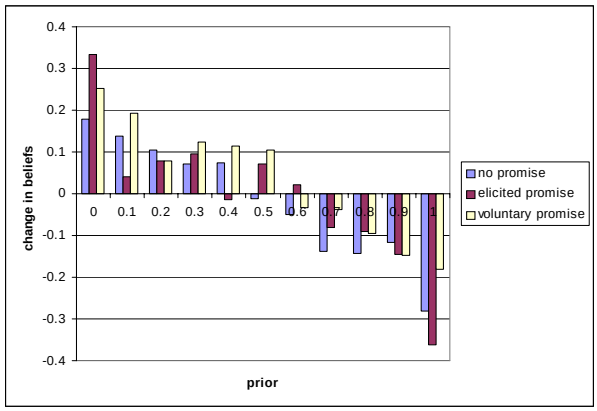

Figure 4: Prior and mean change predictions

beliefs (after communication). Since the equality in (7) holds for each subject, we can aggregate across subjects. We therefore test this hypothesis for each value of interim beliefs, pooling across subjects. Figure 4 shows the mean change in predictions corresponding to each possible value of interim beliefs, and as a function of the type of communication. The pattern is striking: changes are far from being zero on average, at *any* value of interim belief. Final beliefs are systematically higher than interim beliefs when interim beliefs are low and systematically lower when the interim belief is high.[18] It is also noteworthy that subjects with extreme interim beliefs are changing their final beliefs quite dramatically, while Bayesian updating implies that they should change their beliefs very little in response to signals. Nevertheless, our results regarding the voluntary and elicited promises hold even when conditioning on the level of interim beliefs. The positive updates are larger and the negative updates are smaller when subjects see a promise.

The model that best describes our subjects' behavior seems to be one where each subject reports her belief with error. Suppose that the prediction p_{ik} by

[18] A t-test rejects the null hypothesis of the equality of the prior to the average of the posteriors, for each value of interim beliefs (at the 5% significance level).

subject i in instance k is given by

$$p_{ik} = \hat{p}_{ik} + \varepsilon_{ik}, \tag{9}$$

where $\hat{p}_{ik}$ is the "true" belief, from a unimodal distribution centered close to 0.5, and ε_{ik} is an i.i.d. error term. The subject's report equals p_{ik} as long as it lies in the unit interval, and 0 or 1 otherwise. This model can generate the mean reversion in beliefs that we observe, as well as the fact that subjects change their beliefs substantially even when their reported priors are extremal. One interpretation is that subjects find it difficult and costly to uncover their true beliefs, which gives rise to this error. The fact that they are given new information and an opportunity to think again (say after communication) gives rise to a degree of independence in the error term across predictions. This model can potentially explain both our findings. Updating is not systematically of smaller magnitude for extremal reported priors, since these extremal priors are likely to have a larger error term. Mean reversion in beliefs also follows straightforwardly from this model. We leave further exploration of this model for future work.

Our findings relate to an extensive literature on Bayesian updating by experimental subjects. The literature documents several biases (see e.g. Tversky and Kahneman, 1974; Grether, 1980; Gilovich et al., 2002, Charness and Levin, 2005). Some of these biases are potentially present in the current setting, including the confirmatory bias, with subjects recalling or interpreting new evidence as confirming their currently held beliefs (see Rabin and Schrag, 1999; Griffin and Tversky, 1992; and Lord et al., 1979), overconfidence in the precision of own estimates (Lichtenstein et al., 1980; Biais et al., 2005) and over-reaction to recent news (DeBondt and Thaler, 1990; Tversky and Kahneman, 1974). Individuals are often found to underweight new information relative to the prior; however, other individuals exhibit the opposite bias and overweight new information or

show no sizeable bias (El-Gamal and Grether, 1995; Camerer, 1987).

In relation to this literature, our findings are, to our knowledge, somewhat novel, since we reject the martingale property of beliefs, with agents tending to move away systematically from extreme prior beliefs.[19] It is possible that we uncover this since we study updating by experimental subjects in the context of a natural and complex problem, rather than urn-ball experiments, which also tend to have a limited range of prior beliefs. For a Bayesian subject, she needs to think about all possible signals, and her posterior in each of these events. Her prior is a weighted average of these posteriors. It is clear that our subjects do not behave in this way. It is also striking that many subjects have extreme priors such as 0 or 1, which would seem quite irrational, especially in view of their subsequent posteriors. We have suggested that a model of errors in reported beliefs can best describe subject behavior.

6 Subject characteristics

We now turn to the relation between subject characteristics and their beliefs. A first question is what kinds of subjects think that players are trustworthy, and what kinds believe that they are likely to be self-seeking and opportunistic? A second question is whether subjects who are better at predicting differ in a systematic way from others.

6.1 Determinants of beliefs

We first study the determinants of the *level* of beliefs. Our data is eminently suited to study this question since we have a random assignment of players to subjects, and we also have multiple predictions on each player, allowing us to exploit the variation in predictions regarding the same player across subjects. Our second source of information is a subject's predicted average cooperation

[19] The martingale property of *market* expectations has of course been extensively tested in the context of financial markets; however, this may hold even if individual level beliefs do not satisfy the martingale property.

rate or *prior*. Recall that subjects were asked to report the average rate of sharing over all 69 episodes of the show and were rewarded with one extra euro if their prediction was within 5% of the true value.

The prior and the subject mean prediction are correlated, with the coefficient of correlation of 0.38 that is significant at the 1% level. We show the results in Table 6. Columns (1) and (2) are based on the whole sample, columns (3) and (4) are based on the Amsterdam sample only. The striking finding is that the best predictor of a subject's beliefs is his or her own decision in the prisoner's dilemma game. An individual who shares herself is more likely to believe that the game show players will share. This is a striking finding, especially in view of the fact that none of the other subject characteristics is important, once we control for this.[20] That is, subjects who are more cooperative also believe that others are more cooperative. This has a strong resemblance to the finding by Glaeser et al. (2000), that there is a strong correlation between someone's trustworthiness and that person's reported trust in others. They find this result using reported trust levels from survey questions, and we confirm this with the actual predictions of the players' behavior.

A priori there are different explanations for this positive correlation. First, theories of inequity aversion predict that people are more likely to cooperate if they believe that the opponent is likely to cooperate as well (see for instance Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). Second, people may project their preferences onto others, thinking that the average person behaves like them (Messe and Sivacek, 1979). A third explanation is based on a notion of cognitive dissonance (Festinger, 1957) – those who grab may rationalize their behavior by believing that the "whole world behaves like them".[21]

[20] If we do not control for the decisions to share or to donate to the charity, we find a larger positive gender effect (which is significant in the regression based on the same sample as column (1)). However, the gender effect is not as good a predictor as the decision to share in the PD game.

[21] Subjects report their beliefs before they know that they will play the prisoner's dilemma game, so that their beliefs cannot adjust to actual behavior. Nevertheless, it is possible that generic defectors have more pessimistic beliefs than cooperators due to cognitive dissonance.

Table 6 - Subject characteristics, prior and actual predictions

	All sample				Amsterdam sample			
	Predictions		Prior		Predictions		Prior	
	(1)		(2)		(3)		(4)	
Female	.023	(.024)	.008	(.036)	-.013	(.047)	.002	(.063)
Age	.000	(.002)	.00004	(.00001)**	.000	(.006)	.004	(.009)
Number of siblings	.010	(.011)	.001	(.016)	.005	(.020)	-.008	(.026)
Psychology student	-.029	(.033)	-.020	(.050)	-.044	(.040)	.051	(.074)
Other studies	-.002	(.024)	.035	(.037)	-.004	(.040)	.075	(.054)
Employed	.044	(.054)	.102	(.066)				
Shares in PD game	.050	(.023)**	.114	(.034)***	.063	(.039)	.065	(.053)
IQ test score					-.002	(.010)	-.025	(.013)*
Charity donation					.147	(.410)	.485	(.550)
N. subjects	168		168		80		80	
R-squared	.19		.13		.26		.21	

Standard errors are reported in parentheses, *, ** and *** correspond to 10%, 5% and 1% significance levels respectively

6.2 Accuracy of predictions

The second question is whether subjects who are better at predicting differ in a systematic way from others. For example, identifying cooperators is especially important for those who intend to cooperate themselves. We investigate whether the earnings per prediction, E_{ij}, are systematically correlated with the characteristics of the subjects. We estimate the following expression:

$$E_{ij} = \delta_j + \beta' X_i + \alpha_i + \varepsilon_{ij},$$

where δ_j is a player fixed effect and α_i is a subject random specific effect.

Table 7 reports the results. We find some evidence that women are better at predicting than men. Women are also substantially more cooperative in our sample (82% against 46% in the Amsterdam sample, 60% against 23% in the Utrecht sample[22]), so these results suggest that women are also somewhat better at identifying cooperators. Next to that, economics and psychology students perform slightly worse on average than the others, and we find some correlation between risk aversion and quality of predictions. We find no correlation however

[22] In some other experiments women are also more cooperative, although the evidence is mixed (see Eckel and Grossman, 1999).

Table 7: Determinants of the quality of predictions

Dependent variable	Earnings per prediction (OLS estimates)			
	All sample		Amsterdam sample	
	(1)		(2)	
Female	.028	(.033)	.082	(.044)*
Age	-.005	(.003)	-.001	(.006)
Number of siblings	.019	(.014)	.039	(.020)
Economics student				
Psychology student	.024	(.036)	.027	(.048)
Other type of studies	.039	(.033)	.115	(.044)**
Employed	.129	(.057)**		
Shares in PD game	.002	(.030)	.009	(.039)
Prior	-.017	(.075)	.032	(.117)
IQ test score			.011	(.010)
Charity donation			-.405	(.411)
Number of safe choices			.025	(.014)*
Impatient			-.006	(.044)
Player fixed effects	Yes		Yes	
N observations	3680		1896	
N subjects	168		79	
R-squared	.14		.20	

Standard errors are reported in parentheses, *, ** and *** correspond to 10%, 5% and 1% significance levels respectively

between the actual cooperative behavior and the quality of predictions. Those who choose "share" in the Prisoner's dilemma game do not perform better than those who choose "grab". Thus, besides the correlation between gender and earnings, we find very little evidence that cooperators are better at identifying other cooperators. Finally, we find no evidence of a correlation between IQ or time preferences and earnings.

7 Conclusion

We have examined the ability of subjects to predict the behavior of the players of a prisoner's dilemma game. Our key finding is that trustworthiness does appear to be predictable. Subjects' beliefs respond to indicators of cooperative behavior in the correct way. Most importantly, subjects revise their beliefs

upwards in response to a promise that is volunteered by a player, but not in response to a promise that arises due to an explicit question. This suggests that our subjects understand that lies that are volunteered are psychologically more costly. Subjects are also able to distinguish truth from lies when a promise is made in response to an explicit (and unexpected question), in line with the involuntary truth telling hypothesis. Overall, our untrained subjects assign a 7% higher probability of cooperation to cooperators as compared to defectors, suggesting that opportunism is indeed detectable even by naive subjects.

By eliciting subject beliefs at two different stages on the show, we are also able to study the updating process. We find mean reversion of extreme beliefs, in a way that is inconsistent with the martingale property of beliefs. Thus our paper is a contribution to the literature on Bayesian and non-Bayesian updating, in the context of a natural and complex problem.

8 Appendix: Instructions

In the Utrecht sessions, subjects watched the show right up to the point where the lead player chooses the other finalist (see Fig. 2). We asked them to predict the selection decision, and to predict the sharing probability for every player. Because the opponent of the lead player was still unknown to the subjects, we asked them to predict the sharing probability of the lead player against each of the two possible opponents. We then continued the show up to the point where the two finalists had to make their decision to share or grab and paused for the second time. In the Amsterdam treatment, we only asked subjects to predict which of the two contestants would be selected to play the final round when the video was paused for the first time. The show was paused for an additional time right after the selection decision was made, at which point we asked subjects to predict the sharing probability for every player. Finally, we showed them the communication stage between the two final players and paused the show just before the contestants make their final decision to share or to grab.

8.1 Basic Instructions

(Translated from Dutch.) Welcome! The experiment lasts for about 90 minutes and consists of several parts. During the experiment you earn points that are worth money. The exact amount you earn depends on your score and can go up to about €20. None of the other players will know what you earn and all your answers will be treated confidentially.

How you earn money During the first part of the experiment you will see fragments of a television game show. You will be asked to predict choices of contestants. The more accurate your predictions are, the higher your score and the more money you earn. Only your own choices determine your score and not the choices of other participants.

The TV show The game show starts with 5 candidates. Each round, the candidates have to answer trivia questions. Their score depends on the number of questions answered correctly. At the end of the round, one player is eliminated by the highest scoring player. After three rounds, there are three candidates left. At that point, the highest scoring player can decide who to take with him or her into the final. The candidate who chooses is guaranteed to go to the final.

In the final, the scores of both candidates are added. This is the amount of money they will be playing for. Both players simultaneously decide whether to share or shaft. There are three possible situations.

1. They both share. In this case, they both get half of the amount of money.
2. One candidate shares and the other does not share. In this case, the one who does not share gets the whole amount. The candidate who shares gets nothing.
3. They both do not share. In this case, nobody wins any money.

Before making their decision, they have the opportunity to communicate.

(An example was included.)

Instructions You will see 6 shortened episodes. In 2 episodes you'll see one round of trivia questions, in the other 4 we skip all trivia rounds. We start by showing the beginning where the 3 relevant candidates introduce themselves.

The show is paused at the moment 3 candidates are left, and the candidate with the highest score decides who to select to play the final with. At that point, we will ask you to predict the following: Which candidate will be selected to be taken into the final?

After you made your predictions we show you the rest of the episode and pause again when the candidates show their intentions to the viewers at home

(their choice is hidden for you) and when the finalists make their definite choices. At that point we ask you again to make a prediction: Will a candidate will share or not. [in Amsterdam we randomized between predicting sharing and predicting grabbing.]

We ask you to indicate probabilities. For instance, we ask you what you think is the probability that Jennifer will share if she ends up in the final. Imagine you think that Jennifer shares with a probability of 20% (probability 0.2), and hence grabs with 80% probability, then you indicate this as follows:

Prediction choice **Jennifer**										
Probability Jennifer shares	0.0	0.1	X	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

After filling in your answer sheet we ask you to put it in the envelope on your table. After you did this you are not permitted to take it out the envelope. Hence, you can not go back to an earlier question.

Your earnings At the end of the experiment, we compare your predictions to the actual outcomes. You score is higher if your predictions are better. The most you can earn per prediction is 2 points and the minimum is 0 points. Every point is worth €0.35. The amount you earn is calculated by the formula below. This formula is chosen in such a way that it is in your interest to report your true beliefs. By reporting any other number than what you truly believe, your expected earnings are decreased. A proof of this can be requested at the end of the experiment.

Your score depends partly on your choices in other parts of the experiment. Instructions for other parts follow later. Your score in this part does not affect your score in the other parts.

Questions? If you have any questions, please raise your hand and wait until somebody comes to you.

Formula of your score Suppose you reported a probability p that Jennifer will share. If she shares, the score for your prediction is:

$$2 - 2(1-p)^2.$$

If Jennifer decides not to share, the score for your prediction is:

$$2 - 2p^2.$$

Suppose you believe Jennifer shares with probability q. Your expected score by reporting p is:

$$2 - 2q(1-p)^2 - 2(1-q)p^2.$$

You can verify that your expected score is maximized by reporting your true beliefs, i.e. $p = q$.

8.2 Further details of treatments

All subjects were given the above instructions, with minor adjustments. The subjects in Utrecht saw four shows in total, of which 2 complete episodes and 2 shortened episodes. The shortened episodes did not include the trivia rounds, but did include the beginning where candidates introduced themselves. In total we used six shows, stratified by gender composition, percentage sharing, stakes, and percentage of promises and sharing. The order and length (long/short) were randomized among groups. The videos were paused at the moment the candidate had to select one of the other candidates for the final, and at the moment that candidates had to make sharing decisions.

The subjects in Amsterdam saw shortened episodes. Two episodes included the third round of trivia questions. The other four episodes did not include trivia rounds. In addition, subjects only saw written transcripts of the communication of two shows. We showed all remaining episodes. Episodes were randomized among groups. The videos were paused at the moment the candidate had to select one of the other candidates for the final, at the point where a candidate was selected but before they communicated, and at the moment that candidates had to make sharing decisions.

In Utrecht all students were paid the week following the experiments. in Amsterdam students were given the choice to collect their earnings in the week following the experiments, or one month later at a 10% premium. We used the choices of the students to classify them as patient or impatient. Students collecting their earnings early were classified as impatient.

The Amsterdam sessions had two additional tasks. First, we interrupted the video watching after three episodes to ask subjects to do a short cognitive ability test, taken from Frederick (2005) but with four additional questions of similar nature that were kindly provided to us by Shane Frederick. Second, after all the episodes were shown, we asked them to fill-in a questionnaire related to risk preferences. We asked them to choose between a fixed amount (€2) and a lottery (with a 50% chance of earning nothing and a 50% of earning 3, 3.50, 4, 4.50, 5, 5.50, 6, 6.50 respectively; half of the sessions had the reverse ordering of lotteries).

We ended with a questionnaire about their personal background. We also elicited their prior on the probability of sharing, and their own choice in a prisoner's dilemma (the latter was only played for money in Amsterdam). The corresponding questions are:

6. Taken over all episodes (69 in total), what do you think is the percentage of candidates that shares? (with this question you earn €1 in case your answer is within 5 percentage points of the true percentage).

8. We now ask you to play the final of the game yourself. You have to indicate if you want to share or if you do not want to share. Afterwards, we randomly choose two participants of all sessions and their choices are matched. These two participants play for €200,-. The game is played in the same way as in the TV show. So if you both share, both get €100. If one shares and the other does not share, the one who does not shares gets all, so €200. if nobody shares, nobody receives anything.

References

[1] Anderson, D. (1999), The aggregate burden of crime, *Journal of Law and Economics* 42(2), 611–642.

[2] Belot, M., V. Bhaskar and J. van de Ven (2006), A Public Dilemma: Cooperation with Large Stakes and a Large Audience, Working Paper 221, ELSE, University College London.

[3] Biais, B., D. Hilton, K. Mazurier, and S. Pouget (2005), Judgmental Overconfidence, Self-Monitoring and Trading Performance in an Experimental Financial Market, *Review of Economic Studies* 72, 287-312.

[4] Blanco, M., D. Engelmann and H. Normann (2006), A Within-Subject Analysis of Other-Regarding Preferences, mimeo.

[5] Bolton, G., and A. Ockenfels (2000), ERC: A Theory of Equity, Reciprocity, and Competition, *American Economic Review* 90(1), 166-193.

[6] Brosig, J. (2002), Identifying cooperative behavior: Some experimental results in a prisoner's dilemma game, *Journal of Economic Behavior & Organization* 47(3), 275-290.

[7] Camerer, C. (1987), Do Biases in Probability Judgment Matter in Markets? Experimental Evidence, *American Economic Review* 77(5), 981-997.

[8] Charness, G. and M. Dufwenberg (2007), Broken Promises: An Experiment, mimeo.

[9] Charness, G., and D. Levin (2005), When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect, *American Economic Review* 95(4), 1300-1310.

[10] DeBondt, W., and R. Thaler (1990), Do security analysts overreact?, *American Economic Review* 80(2), 52–57.

[11] Dekel, E., J. Ely and O. Yilnakaya (2007), The Evolution of Preferences, *Review of Economic Studies* 74(3), 685-704.

[12] DePaulo, B.M., J. Stone and D. Lassiter (1985), Deceiving and detecting deceit, In: B. Schenkler (Ed.), *The self and social life*, New York: McGraw-Hill.

[13] DePaulo, B.M., and H.S. Friedman (1998), Nonverbal communication, In: D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of Social Psychology* (4th ed., Vol. 2), New York: McGraw–Hill.

[14] Eckel, C. and P. Grossman, (1999), Differences in Economic Decisions of Men and Women, in C. Plott and V. Smith (eds.) *Handbook of Experimental Results,* Amsterdam: Elsevier.

[15] Ekman, P. (1985), *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*, W. W. Norton.

[16] Ekman, P., and W. Friesen (1974), Detecting deception from body or face, *Journal of Personality and Social Psychology* 29, 288-298.

[17] Ekman, P., M. O'Sullivan, and M. Frank (1999), A few can catch a liar, *Psychological Science* 10, 263-266.

[18] El-Gamal, M.A., and D.M. Grether (1995), Are People Bayesian? Uncovering Behavioral Strategies, *Journal of the American Statistical Association* 90(432), 1137-1145.

[19] Fehr, E., and K. Schmidt (1999), A Theory of Fairness, Competition, and Cooperation, *Quarterly Journal of Economics* 114(3), 817-868.

[20] Festinger, L. (1957), *A theory of cognitive dissonance*, Row, Peterson and Company.

[21] Frank, M., and P. Ekman (1997), The Ability to Detect Deceit Generalizes Across DifferentTypes of High-Stake Lies, *Journal of Personality and Social Psychology* 72(6), 1429-1439.

[22] Frank, R. (1988), *Passions within Reason: The Strategic Role of the Emotions,* New York: W.W. Norton & Company.

[23] Frank, R., T. Gilovich and D. Regan (1993), The evolution of One-Shot Cooperation: An experiment, *Ethology and Sociobiology* 14, 247-256.

[24] Frederick, S. (2005), Cognitive Reflection and Decision Making, *Journal of Economic Perspectives* 19(4), 24-42.

[25] Gilovich, T., D. Griffin, and D. Kahneman (Eds.) (2002), *Heuristics and Biases*, Cambridge University Press, New York.

[26] Glaeser, E., D. Laibson, J. Scheinkman, and C. Souter (2000), Measuring Trust, *Quarterly Journal of Economics* 115, 811-846.

[27] Gneezy, U. (2005), Deception: The Role of Consequences, *American Economic Review* 90(1), 384-394.

[28] Griffin, D., and A. Tversky (1992), The Weighing of Evidence and the Determinants of Confidence, *Cognitive Psychology* 24, 411-435.

[29] Grether, D.M. (1980), Bayes Rule as a Descriptive Model: The Representativeness Heuristic, *Quarterly Journal of Economics* 95(3), 537-557.

[30] Guth, W., and M. Yaari (1992), Explaining reciprocal behavior in simple strategic games: an evolutionary approach, Chapter 2 in: Witt, U. (Ed.)., *Explaining Process and Change: Approaches to Evolutionary Economics,* University of Michigan Press, Ann Arbor, MI, pp. 23–34.

[31] Knack, S., and P. Keefer (1997), Does social capital have an economic payoff? A cross-country investigation, *Quarterly Journal of Economics* 112, 1251-1288.

[32] Kraut, R. E., and D. Poe (1980), Behavioral roots of person perception: The deception judgments of customs Inspectors and laymen, *Journal of Personality and Social Psychology* 39, 784-798.

[33] La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny (1997), Trust in Large Organizations, *American Economic Review Papers and Proceedings* 87, 333-338.

[34] Laband, D., and J. Sophocleus (1992), An estimate of resource expenditures on transfer activity in the United States, *Quarterly Journal of Economics* 107, 959–983.

[35] Lichtenstein, S., B. Fischhoff, and L. Phillips (1982), Calibration of Probabilities: The State of the Art to 1980, in Gilovich, T., D. Griffin, and D. Kahneman (Eds.) (2002), *Heuristics and Biases*, Cambridge University Press, New York.

[36] Lord, C.G., L. Ross and M.R. Lepper (1979), Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence, *Journal of Personality and Social Psychology* 37(11), 2098-2109.

[37] Mann, M., A. Vrij, and R. Bull (2002), Detecting true lies: Police officers' ability to detech suspects' lies, *Journal of Applied Psychology* 89(1), 137-149.

[38] Messé, L. A. and J.M. Sivacek (1979), Predictions of others' responses in a mixed-motive game: self justification or false consensus?, *Journal of Personality and Social Psychology* 37, 602-607.

[39] Ockenfels, A., and R. Selten (2000), An Experiment on the Hypothesis of Involuntary Truth-Signalling in Bargaining, *Games and Economic Behavior* 33(1), 90-116.

[40] Rabin, M., and J. Schrag (1999), First Impressions Matter: A Model of Confirmatory Bias, *Quarterly Journal of Economics* 114(1), 37-82.

[41] Tversky, A., and D. Kahneman (1974), Judgment and uncertainty: Heuristics and biases, *Science* 185, 1124-1131.

[42] Wang, J., M. Spezio, and C. Camerer (2006), Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth-Telling and Deception in Biased Transmission Games, mimeo, Caltech.